Understanding Forensic Digital Imaging

Herbert Blitzer
Karën Stein-Ferguson
Jeffrey Huang
With contributions by Jack Jacobia

Dedication

This book is dedicated to Sara Blitzer who provided support to me to finish my college education in physics and then start my professional career with the Eastman Kodak Company. She did this right after becoming widowed.

Acknowledgments

The authors wish to thank Lauren Marina Cregor, MA, for the initial proofreading of the manuscript and providing a host of helpful suggestions. Also, to William Oliver, MD, for motivating this book by arguing the proposition that experts should be able to explain the basic processes associated with the image processing tools they use.

Forward

In American culture the field of law and forensic science is often dramatized and over simplified through television and media reports. Between "Law & Order" and "CSI" the public is exposed to scenarios of crime scene investigation and prosecution that are hyped on emotion and lacking in scientific foundation. The result is that the general public has a romanticized idea of what truly happens in the justice system and they have unrealistic expectations of what science and the law is actually able to provide to a finder of fact in a legal setting.

Forensic science is where law meets science in a forum where expert witnesses must be prepared to explain and defend their conclusions.

Webster's Dictionary defines "forensic" as: "belonging to, used in, or suitable to courts of judicature or to public discussion and debate." The American Heritage College Dictionary defines "imaging" as: "to translate (photographs or other pictures) by computer into numbers that can be transmitted to and reconverted into pictures by another computer." This book is intended to support initiatives that will allow individuals to be better educated about the science of forensic imaging and the preparation necessary to offer testimony concerning forensic imaging in a legal context. It is important to first remember the forum where the science meets the law, and then that the expert must be able to translate the scientific techniques and principles into a language and conclusion that a lay person will feel they can rely upon. Judges and juries are lay people for the most part and need to be properly educated on what information can be relied upon and what cannot. A true expert is also prepared to explain the limitations of the science without apology or defensiveness so that the judge or jury can decide what weight and credibility should be given to the results.

There are countless stories in the media of individuals who were convicted and imprisoned and advances in forensic science later proved them to be innocent. Literally the results of scientific testimony can be a matter of life and death. However, juries have come to expect forensic science findings will be presented at trials and may automatically judge the case weak if it is not.

Forensic evidence is presented in a Court of law by having an individual qualified as an expert describe the facts available, any quantitative or qualitative measurements made, the application of the science to those facts and measurements, and the conclusions drawn stated as a matter of scientific certainty. This approach to presenting expert testimony is time honored and nothing new. But science is not a static field of expertise, and as it progresses in technique and sophistication, the law and the witnesses presented to explain the science must adjust as well. There are many examples of cases that become battles of the experts. Evidence that might seem compelling and unimpeachable one day may be regarded as out-dated and unreliable the next. An example is fingerprint evidence which is often portrayed on television as almost as finite as DNA. However, the reality of gathering, preserving and interpreting fingerprints is often a matter of controversy.

This book reviews the field of digital imaging and the science behind the more common tools and techniques is revealed. Anyone can snap a picture with a digital camera and produce an image rather easily using the available software. Being able to analyze that image and testify as to the content and whether the image is a "true and accurate depiction of what it is intended to portray" is a whole different responsibility. A number of complex tools must be used to analyze an image and testify that it has not been tampered with or the image distorted in a way that can skew the interpretation of the image. The expert must then be able to explain the basis for selecting the tools that were used, the order in which they were used and why the judge or jury should believe that these tools were the best and most appropriate to use in the analysis in question.

Imagine that you are the member of a jury in a case involving allegations of domestic violence. The prosecutor introduces photographs through an expert that seem to depict serious injuries to the victim of the domestic abuse. The photographs show what appear to be redness, abrasions and possible small lacerations to the victim's face and the prosecutor argues that these are the result of a beating. The evidence seems most compelling. The defense then brings in an expert who testifies that the photographs are, in fact, quite misleading. The defense expert attacks the camera used, the inappropriate settings on the camera at the time the photographs were taken, how a combination of factors has caused exaggerations or artifacts in the images, and the fact that the victim suffers from a severe case of acne so that it is impossible to separate injuries from the skin condition given the photographic tools and techniques that were employed. What appeared to be compelling evidence may be interpreted as an effort on the part of the party offering the evidence to distort the truth.

The case above is a simple example. But the use of forensic imaging is becoming more and more diverse. The areas in which imaging is being used include fingerprints, footwear and tire impressions, ballistics, tool marks, accident scenes, crime scene reconstruction, documentation of wounds or injuries, surveillance videos, and many others. Many of the cameras, scanners, software suites, printers and monitors or projectors are designed primarily for the consumer market or the artistic/commercial market. These tools are adequate when the intent is merely to evoke emotion or even create special effects. Knowledge of the science is not necessary or even considered.

But in forensic science the objective is quite different. The expert needs to be able to state a conclusion and feel confident in convincing the judge or jury that the conclusion is valid. To do this it is necessary to be able to know the major elements of the science behind the tools and explain what was done and why it was done that way. Forensics should be driven by truth seeking, not emotional impact.

As with any field of science, those now preparing to enter the field of forensic science will need to be better prepared and educated than their predecessors. They will also need to keep up with the ever accelerating pace of change. It is hoped this book will assist in supporting the new college curricula and expanding degree programs in the field of forensic science.

Sonia J. Leerkamp, Prosecuting Attorney,
Hamilton County, Indiana

Introduction

There are books that teach digital imaging technique and courses that teach one how to design cameras, computers, software, and other high-tech devices. The former are necessary to actually processing a case, but the content is time sensitive because the specific devices and software packages change frequently. The latter are for engineers that will be designing devices and software for practitioners. This book is positioned between these two approaches. It discusses the science behind the devices and software and helps explain why commercially available items work the way they do and how to best use them to solve problems. It goes further in that it helps the forensic expert equip himself to answer tough questions that might arise regarding why he did what he did and why that is valid.

The scientific basis is several decades old. Sharpening filters, unsharp mask techniques, brightness and contrast adjustment tools, and many other tools are derived from darkroom techniques that were developed, in some cases, over 100 years ago. The mechanism is now digital instead of analog, but the approach is the same. It is not likely that it will change dramatically in the near future; therefore, the material will have a certain degree of durability. There are some new digitally enabled tools that perform actions that are very difficult with analog photography, but the basis for these is not fundamentally new. For example the Fourier transform goes back to the early 1800s. It is just that modern computers make it an easy and fast tool to use.

The first four chapters: Why Take Pictures, Dynamic Range, Light and Lenses, and Photometry are quite general and are the foundation for much of what follows. The next chapter, Setting Exposures puts some of the basics together in ways that apply to photography. Then the chapter, Color Space brings up an old concept that needs to be made digital and is a cornerstone of digital photography and image processing. It is also a key in that it carries the means for the human visual system to utilize photographs. The chapter on Showing Images deals with the basics of how printers and monitors work.

The chapter, Key Photographic Techniques is a sampling of the schemes that photographers have developed since the earliest days of photography, and their use in the digital age is explained. This is followed by a chapter on Image Processing Tools. Only the more common ones are described because there are so many of them. The emphasis is on how they work as opposed to how to work them.

Digital scanners tend to be in the background of digital photography. It is the cameras that get all the attention. Nonetheless, scanners can deliver excellent images in cases where a camera would struggle.

At the heart of any digital device are special electronic circuits and a non-intuitive number system. It is said that people work with groupings of ten (the decimal system) because they have ten fingers. Digital circuits, by comparison are most easily made to deal with groupings of two, so is convenient to have those circuits work in a number system based on groupings of two (the binary system). This chapter, which is not an easy read, will help the practitioner understand what is happening with the mysterious "zeros and ones". This material is a good lead into the chapter on File Formats and Compression. These are separate issues but they tend to be tightly intertwined, and can only be appreciated at the basic level by understanding that they work with binary data.

The next three chapters get into some key equipment issues. The chapter on Sensor Chips describes the basics of how these magnificent devices work. The next chapter on Storage and Media describes the more commonly used devices and how they work. Finally, the chapter on Computing Images describes what happens in the camera to convert an optical image from an exposure into a sensor chip response, and then to an outputted image file.

The chapter on Establishing Quality Requirements brings together material from all of the preceding chapters and explains how one can determine what a lab might want to do for different disciplines. It goes on to provide basic calculations and methods that can be used.

The Scientific Working Group on Imaging Technology (SWGIT) has been developing and publishing guidelines for the use of imaging technology in forensic applications for the past decade. This chapter gives a summary of some of their key issues. The guidelines themselves are constantly being updated and are available on the Internet, so they are not reproduced here.

With the science, quality requirements and guidelines in hand, one is ready to review the relationship between Digital Images and Investigations. This is followed by a chapter on Getting Digital Images Admitted as Evidence at Trial. Included are elements from the rules of evidence and analyses of several key cases. The applicable Federal Rules of Evidence are in the Appendix.

As should be apparent from the descriptions of the book contents, the material has several convolutions. This means that topics will come up with some degree of repetition and in various combinations. In many of these cases, material that was discussed earlier is refreshed in a later context. The hope is that this will minimize excessive page flipping.

Many of the chapters have either thought provoking questions or exercises attached. These help drive home the contents of the preceding chapter. Some of the exercises require downloading items from the book's website.

CHAPTER 1

Introduction to Forensic Use of Digital Imaging

WHY TAKE PICTURES?

Taking pictures is such a normal thing to do that we rarely think about why we are doing it. This is especially true today when cameras are so ubiquitous and easy to use that you can take photos with your cell phone. You don't have to buy film or have it processed, and you might never print some photos or even show your photos to anyone. So why do it? In one of their most effective advertising campaigns, the Eastman Kodak Company addressed the idea of converting special events into memories, and called those situations "Kodak moments." The most common reason for taking pictures is to jog our memories at some later time and bring back the feelings of that moment. Humans are very good at using these visual clues to resurrect the whole set of feelings and understandings that the photo preserved. This means that the photographer does not really have to be particularly skilled to get photos that will serve the purpose. The amateur photography industry is predicated on these simple facts:

- Photos are very good at bringing to mind whole scenarios from the past
- People appreciate reliving certain moments
- Photos are easy to take
- The cost is very reasonable

This has been the case since the 1880s. Prior to that, in the 1860s, photos were being taken, but the complex nature of the technology at that time limited its use to professional photographers. Photos from the civil war in the United States are still compelling to all who see them, but only Matthew Brady and his colleagues could take pictures back then.

But what about before that time: What were the precursors to photography? Drawings and paintings are the obvious responses. These go back to the Stone Age. Unfortunately they require some skill to produce, and if the individual is not so skilled, an artist has to be hired, so the cost is not right for everyone. Most people can make sketches, though, and in many instances that had to suffice. Some of these were no doubt quite rough indeed. Another approach to preserving memories was with verbal descriptions. These could be told around a campfire and easily embellished over time to suit the purposes of each story teller. Adding melody made it easier to remember the words and captured additional feelings. When writing came into being, the oral history could be rendered as a written history. These were effective, could be extended over long time periods and distances, and although embellishment was possible, it was not quite as easy as with the oral version. Drawings and pictures could be added easily, and decorations could be put on the pages to reinforce the importance of the material. All these memory-jogging techniques continue to this day. One interesting aspect of the memory jogger is that it generally requires that the reader have a memory to jog. That is, he was there at the time of the original event, can envision a reasonable semblance of that situation, or has heard or seen the story so often that he has a mental image of it even though he was never there.

In the world of forensics, some of the factors change. First of all, the memory-jogging mission applies only to the people who were there at the time. For all others, the issue is communication. In this situation, the person who was there at the crime scene, the accident scene, or the disaster scene is trying to convey to others what the scene was like, what was there at the time, where those things were in relation to each other, and what condition the items were in at the time. The simple internal, emotional glow of the memory jogger (assuming a happy event) gives way to a more matter-of-fact communication. The photographer, or someone else who was at the scene, will be asked to confirm that the photo is a fair and accurate representation of what was there at the time. This process is sometimes called *visual verification*. The people who were there can say, in essence, "I was there and it looked like what you see in the photo." One could use a sketch in such situations, or the description could be simply verbal (written in a report or transcript) or oral (during testimony). The photo however will contain much more detail. And in most situations, time is of the essence; creating a complete and meticulous written listing of what was there and where it was would be difficult, to say the least. Moreover, it would not convey the ambiance of the situation nearly as well as a photo. Without a photograph, the effect of the lighting will be gone, the comprehension of the level of general orderliness (or confusion) will be lost, and the character of any decoration will vanish. Just imagine a person trying to give an oral description of a tire track impression in sufficient detail so as to allow a determination of whether a confiscated tire made a particular track. The photo conveys the gestalt of the setting, not just a few details.

A photo can convey a comprehensive impression of an environment, and since much will depend upon doing this fairly and accurately, the photographer and subsequent image preparer must do their work with more skill than the average amateur to avoid the bias of the freelance storyteller. The photos must be exposed properly to give the viewer a clear impression of what the scene was like at the time. They must show both the relationships among objects as well as detail in key areas. This is usually accomplished by taking establishment shots from some distance away, medium shots to juxtapose selected items accurately, and close-ups to show important details. Finally, it is important to avoid bias.

Freelance photographers are often out to tell a story as opposed to presenting a balanced set of facts. As a result they will carefully compose photos to do just that. For example, if the story involves enforced separations, they will look for some fencing and then position a subject in front of that fence to help the storyline even if the fence in the photo has nothing to do with the separations. If they are seeking to express slovenliness, they may take photos in a workshop or laundry room at some inopportune time. In general, they have a preplanned story to tell and are looking for ways to convey that message. In forensic assignments, the story is probably not known at the time the photos are taken, and in fact, the photos should be able to play an important part in determining what the true story is. But it must be a fair and accurate story. Then, later, they can be used to help tell that story to a jury or judge.

In the typical forensic photography assignment, the timeline is an important issue. The first representatives of authority on the scene are normally patrol officers. They ascertain the nature of the situation, care for any injured people, and at the same time, protect the area from contamination and change. The technicians, including the photographer(s), will be next on the scene. They have limited time to document the setting as it was found, and to collect samples and items that could be useful in understanding what happened. As they do their work, the scene will start to undergo change, and as they complete their assignment, the rate of change will accelerate. There is no going back. They must get it right the first time. While they are working the crime scene, other investigators are starting to question witnesses. The story will begin to unfold. And later, after a lot of detective work, the story of the situation will start to become clear. This means that the photographer(s) had to do their work without knowing the story their work eventually would help to tell. In most jurisdictions, all the photos taken by the police or crime lab may have to be given to the defense team. So any attempts to bias the story using photos taken before the whole story is known could lead to extremely embarrassing outcomes and the release of a potentially dangerous defendant. Fairness is required.

The most common purpose for photos is to revive memories, the second is to communicate, and the third is to provide a base for measurements. If the purpose for the photos is to recall memories, no special care is required in taking the photos. If the purpose is to tell a story, a sequence of photos will be needed, and it must be possible for viewers of the images to make the connections among the various shots. If the images will be used for making measurements, great care must be taken to ensure that the intended measurements will be valid. The particulars will vary with the anticipated analytical purposes. In many instances, special analytical tools are used to extract information from photographs. Some tools extract dimensions or colors that are attributable to the item that was photographed. More recently, sets of photos have been used to create three-dimensional renditions of objects. In these situations, great care must be taken to ensure that when the photo(s) was taken close attention was paid to the intended measurement process that would follow. A significant amount of image processing, sometimes using complex tools in complicated combinations, might be used to prepare the image prior to measurement. Some of those processing tools might introduce distortions that could make the measurements difficult or inaccurate if not properly applied. In a number of image measurement situations, the image that actually is measured may not be visually verifiable. This arises when the object is not visible to the human eye, and therefore, no one actually could have seen the result prior to processing.

In these situations, the person who analyzed the image has to be able to show that the end result was properly and scientifically extracted from an original photo and that the original photo was a properly and scientifically constructed representation of the original scene or object.

The subsequent chapters of this book explain the basics of the science supporting the most frequently used tools and techniques in forensic photography. The objective is to make the analyst aware of the principles upon which the tools are based, the limitations associated with those tools, and to some degree, why the tools and techniques are designed the way they are. The chapters at the end of the book describe the applicable law and thereby provide guidance to the analyst as needed as he prepares to deliver testimony regarding the work done and the conclusions drawn.

PHOTOGRAPHY AS A SURROGATE

As indicated, photography serves as a surrogate for actually being at the scene. This is generally taken for granted, but in fact a lot of careful design work was required to make the equipment and software suitable for the task. The photographic system employed must capture the *optical* information from a scene; in most cases this is the visual information. This is the information that a person at the scene would be able to glean visually.[1] The photographic

[1]In certain situations the object is being illuminated and photographed using light that is outside the range of normal human vision, in which case other precautions must be taken to validate that the image that is created truly and accurately renders what it purports to show. This is often referred to as Alternative Light Source (ALS) photography. Extreme examples of images from nonviewable originals include x-rays, sonograms, PET scans, and nuclear autoradiographs.

system must then process that information and render it in such a way that a person looking at the image will recognize what he or she is viewing. That is, they can look beyond the photograph and form a mental image of what the original setting was like.

Humans see color by virtue of sensor organs in their eyes called *cones*. These are in the retina on the back, internal wall of the eye. There are three kinds of cones. The first type is responsive to shorter wavelengths in the blue portion of the spectrum; the second is responsive to midrange wavelengths in the green/yellow portion of the spectrum; and the third is sensitive to longer wavelengths reaching out into the red portion of the spectrum. In addition to cones, there are sensors called *rods*. These have broad sensitivity with a peak in the green/yellow range and are used for seeing in darker settings. The rods and cones actually move back and forth depending on the light level. Outdoors at night we use primarily rod vision and during the day, we use primarily cone vision. Since the three types of cones are sensitive to different portions of the visual spectrum, they respond differently to different colors in the original scene and we are able to determine that color by combining those responses. Rods have a broad response, covering the full spectrum, and so respond the same no matter what the color of the object in the scene. We cannot distinguish colors with pure rod vision (Fig. 1.1).

It should be noted that color is a mental construct. The light that we see as yellow is not necessarily a light with a particular wavelength. Roughly equal responses by the red and green cones, and none by the blue cones, will evoke the color yellow. That could be done with some red and some green light, or just a single yellow source. Wavelengths do not have "colors"—humans do.

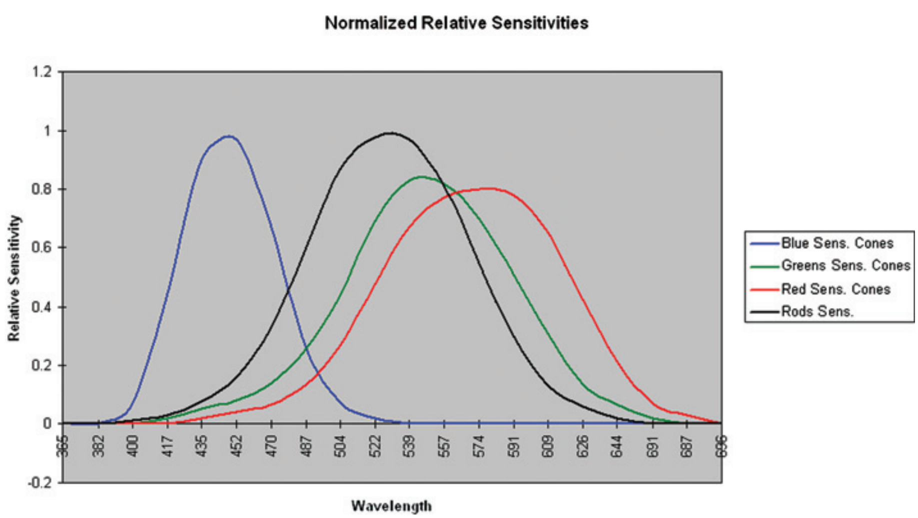


FIGURE 1.1 *Human Eye Sensitivity. The sensitivities of the red, green, and blue sensitive cones in the human eye are shown normalized to the areas under the curves being equal to one. The sensitivity of the rods is shown with its peak sensitivity set to one.*

A photographic system must be able to respond to scene coloration so that it captures information in a way that can be used to construct an image with proper colorization so that a human can recognize the contents.

Once the image information is captured, it must be processed so that it can drive a printer or display device to present a human viewable image. It is easiest to understand the process by skipping to the viewing of the image.

Humans see in their brains, specifically in the occipital lobes, which are located in the back of the head. The eyes capture information and feed it into the optic nerves, which connect into the occipital lobes. The rods and cones in the eye gather the raw data and the visual system starts to process that data in the ganglion cells in the retina. Light levels, primitive shapes, and early blending of color-response start there and move on into the optic nerve. The partially processed information arrives in the central brain[2] where it is assigned meaning and receives detailed analysis. The brain-resident, ephemeral image is held there pending updates from the early parts of the system. It is postulated that the early processing of visual information allows for quick response to emergency situations, such as avoiding predators or responding to prey.

As a person continues to look at a scene, the eyes automatically dart around the area capturing slightly different views. At each point, the eye refocuses and adjusts for light level. The upgraded information is passed along the optic nerve to the brain where the slightly different views are combined and details are filled in. The brain identifies elements in the scene; once this is done, a mentally complete rendition is available even if some of the details are still lacking. The result is that almost everything seems to be in focus, the extremes in light levels are taken into account, and the images from the two eyes are combined mentally to create a three-dimensional view. It is quite a remarkable system![3] There is no photographic system that can do all this, not even close. Humans see the elements of a scene as identifiable objects and ascribe details to them. Mechanical systems see primarily the details and do not see the objects. New software is being included in digital cameras to start to process more information internally, as the eye and optic nerve do. And workers in the field of biometrics are attempting to use computers to process images and determine certain basic information about objects in a scene. But these, though mathematically complex, are primitive by comparison to human vision. A person can look into the street and see a blue car, and know that it is the same blue car even if the shadow of a cloud passes over it. Computers struggle with this.

In the photographic process the image that is presented to the viewer must be recognizable. The basic shapes will be determined largely by the

[2] The retina, optic nerve, and sometimes even the whole eye often are considered part of the brain.

[3] There are animals such as birds and squid that have even more remarkable visual systems.

rods and the creation of shape primitives; coloration will be determined from the responses by the cones. Finally the whole visual system has a remarkable ability to interpret the flat representation as a surrogate and create a full version of the original scene. If the intent is to make a color print, the printer must put in place colorants that will stimulate the red, green, and blue sensitive cones in the correct relative amounts. Likewise an image on a screen must also evoke the same type of response, even though the print does this with a set of colored dyes and the screen device does this with a different set of lights. If this is not done correctly, the viewer will infer the wrong colors and the result can be extremely ineffective as a surrogate (Fig. 1.2).

The input is defined by both the original scene and the device being used for image capture (camera, scanner, etc.). The output is defined by the image-rendering device (printer, display, etc.) and the human visual system. So, the processing requirements are defined by those steps necessary to convert the inputs available to the outputs. It turns out that there are many

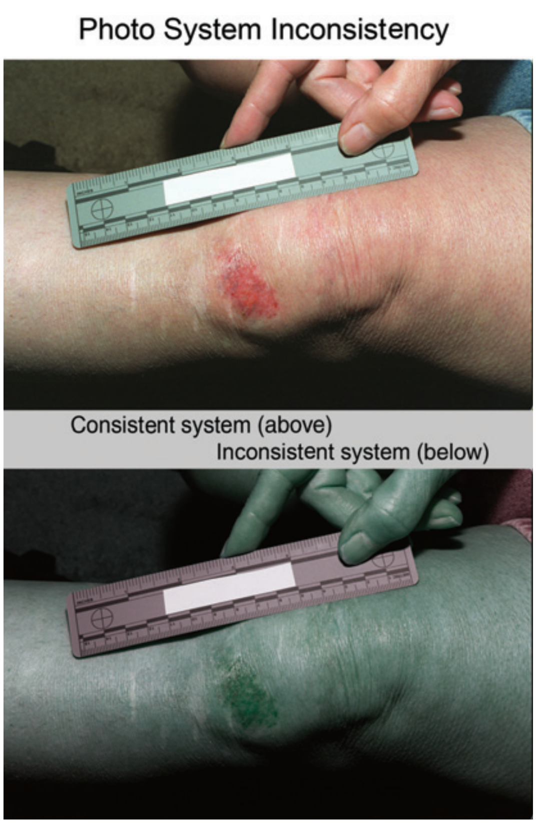


FIGURE 1.2 *Photo System Inconsistency. The figure shows two renditions of the same original photograph. The one on the top image was rendered with a color set that complements the photographic technology color set. The lower image was rendered using a different color set. The lower image is not interpretable.*

steps to the process and they are quite complex. In later chapters the most commonly used of these will be described.

An archival record of the image is an important product of a forensic photographic system. A faithful reproduction of the input must be available for some time into the future in order to facilitate a review of the processes employed, the results obtained, and the ability to use new tools to extract more information from old images. There are three key factors to consider: the storage medium, the image file format, and the process for updating the archive. The image should be recorded on a medium that is known to be reliable and relatively long lasting, and the file should be kept from those who do not require access, and it should be refreshed in a timely manner. The file format should be an open standard in common use. Compression should be avoided since it multiplies the damage due to any lost bits of information. The concept of long lasting is an important issue. It means that the medium and file format used will last until that type of media and image format start to become obsolete. Prior to obsolescence the records in the archive will have to be rerecorded in the new ways. The archive must be actively managed. In forensic applications the duration of an archive can be very long: approaching a century.

Modern photography has gotten to the point where it:

- Is quite easy to use because of several automated features
- Can be arbitrarily accurate
- Can take photos of things that cannot be seen by humans

Also, there is a wide range of analytical tools that aid in the extraction of information from images. In forensic applications, it is important for the examiner not to let the automatic adjustments take free reign and to use the analytical tools with proper care. Otherwise, the result can be misleading. The range of assignments is so great that there is no single path that will work in all situations. The examiner must develop and implement a strategy for each image. This requires that the examiner using the newer technology understand the tools and techniques at a level that is deeper than just how to push the buttons. This book will describe the key underpinnings of several automated features and analytical tools to help practitioners become savvy in their trade.

SOME HISTORY OF FORENSIC PHOTOGRAPHY

Prior to 1880, photographers coated light-sensitive materials onto glass plates just before taking photos, and then processed them immediately afterward, while they were still wet. The major invention that changed the photographic world came when George Eastman learned how to make dry plates and built a factory to coat them. Later came the development of flexible film materials. The films were coated in a factory and then the images were processed in a central laboratory long after the exposures were made.

When this happened, it became practical to take photos at crime scenes. As photographic technology advanced, its use in forensic applications expanded as well. For example, photographers learned how to use contrast-enhancing filters and how to take photos with infrared and ultraviolet light. More recently, video photography has become widespread in surveillance applications, and more and more police cars are being outfitted with cameras to document the behavior of both the police officer and suspect, and to help with officer safety. And, of course, since the mid-1990s, law enforcement has been making use of digital photography.

Historically, the use of photography reaches back to before the invention of silver halide (film) photography. Earliest uses of photography in law enforcement involved Daguerreotype photography, a precursor to silver halide film technology, in Paris in 1841 and in Belgium in 1843. These included the recording of what today we would call mug shots and fingerprint photos.

Not long after that, in 1851, came the first documented case of a manipulated image. Reverend Levi Hill claimed to have developed a way to capture Daguerreotypes in color. He presented an image to show the result. Marcus A. Root studied the image and found that it was colored with fine, dry, colored powders. Clearly, Hill had colored the image by hand. So, image manipulation is not a new phenomenon; it is just that the new digital technology has made it much easier to do. The ability to detect manipulated images is a skill that is still in demand, however when the changes are made by an expert, recognizing these altered images is very hard to do.

Since the mid-1990s the issue of acceptance of digital images has grown in importance. The obvious concern is that digital images are easily manipulated. Thus the party offering the image as evidence must be able to satisfactorily speak to the provenance of the image being offered. This issue has been addressed by special groups formed in a number of countries. In the United States, the group is the Scientific Working Group on Imaging Technology (SWGIT), and in Great Britain it is the Police Scientific Development Branch (PSDB). There are also groups in a number of other countries, including, but not limited to Canada, the Netherlands, Germany, and Australia. These groups have worked both alone and in concert and most of the major issues have been addressed. Most of the conclusions and recommendations are very similar. In this book, the SWGIT guidelines are reviewed in Chapter 18. The main thing to know at this point is that in the United States, no photo has ever been kept out of a trial simply because it was digital. Any problems that have arisen involved the processing of the image and the conclusions drawn from them. These issues are addressed in Chapter 20.

FILM VERSUS DIGITAL PHOTOGRAPHY

With film photography, the film that is in the camera is sensitive to light over its entire surface. The light coming through the lens impinges on that surface and activates silver halide crystals in the sensitive layer; the more light, the more activation. The array of activated sites in the film is referred to as a *latent image*. When the film is processed, the silver halide crystals with active sites are converted from silver halide to silver. In color films, colored dye is formed at the sites as a byproduct of the silver conversion. The result is a film substrate with a coating on its surface containing dye in areas that were exposed to light; the more light, the more dye. This is a color negative. To make a print, light is sent though the negative and focused by a lens onto a paper coated with material that is very similar to the original film. In areas of high exposure, large amounts of dye are formed, and in areas of low exposure, small amounts are formed. Since the overall process involves a two-stage tone reversal, the print is light in areas that were originally light and dark in areas that were originally dark. In other words, the print is a positive comprised of two cascaded negative processes. The negative is a physical record of the original scene and generally is considered to be the *original*.[4]

In the case of digital photography, there is no film. Instead there is an integrated circuit sensor chip. This chip has a very large number of very small surface spots in a regular array. Each surface spot is sensitive to light and they are all independent of each other in their response to incoming light. Often these surface spots are referred to as *pixels* (picture elements). Since each pixel has a defined location on the sensor chip surface, and each has an independent electronic response to the incoming light, the array of electronic responses is a record of the original scene, not unlike the latent image phase of a film record. The next step involves converting each of the electronic responses into a number that represents the amount of light that fell on each pixel. The result is a string of numbers. Each has a pair of location numbers (from the initial sensor chip) and a light level number. The result is that the initial image in digital photography is nothing more than a long string of numbers. Until the numbers are fixed onto a physical medium, there is no tangible record of the image. SWGIT refers to this ephemeral image as a *primary image*, and the first record of that onto a physical medium that will be kept is called the *original image*. Modern cameras also attach a lot of additional information to the image file, and this additional information is called *metadata*. Scanners do not necessarily attach metadata, but they, too, create a string of numbers as the primary image, and until the string is fixed onto a physical medium, there is no original. This is because the primary image will be erased in due course and the surviving version of the image will be the fixed version. It has the same string of numbers as the primary, but it is fixed to a physical medium.

We often think of digital photography as distinct from film-based photography, but that separation is unrealistic. In 2000, Dr. Robert Davis, a consultant

[4] The Federal Rules of Evidence have been interpreted also to call all prints made from the negative "originals" as well. Not good science, but legal.

and educator in Dallas, Texas, demonstrated to the SWGIT that with high-quality digital devices, any image can be corrupted. For his demonstration, he took a number of photographs of a water tower with writing on its face. The pictures were taken with KODAK EKTACHROME film. The images were scanned and converted to digital form using a high-quality film scanner. He then edited the images to remove the writing from the water tower. He wrote the images to the same type of film using a film writer and had the slides mounted. Finally, he sent both sets of slides to former colleagues at the research laboratories of the Eastman Kodak Company. He asked them if they could tell which slides were the originals. They could not. This should not be surprising to us today. We have all seen movies like *Jurassic Park* and *Star Wars*, which have computer-generated characters and creatures mixed in with live actors, and it all looks perfectly real. The same is true of most of the advertisements we see on TV or in magazines. The point of this is that in today's world, the technology used to capture an image or the medium on which an image resides is no guarantee, all by itself, as to the legitimacy of the image. Any image can be altered. Practitioners of forensic imaging must take care in their processes to ensure legitimacy. In a private conversation with a former governor of Indiana (Robert Orr), Randall Shepard, Chief Justice of the Indiana Supreme Court, said, in effect, that ultimately, it comes down to the veracity of the witness, testifying under fear of perjury, that supports the legitimacy of an image. An expert witness must be able to explain his actions to a jury and defend those actions in a cross examination. As jurors become more familiar with the new technology, they will demand better explanations, and as trial lawyers become more aware of the potential for error, the cross examinations will become more pointed and difficult.

QUESTIONS

1. Why does SWGIT differentiate between a primary image and an original image?
2. What are the three main reasons for taking photos, and how does each fit into forensic photography applications?
3. Describe film and digital image originals in terms of their physical condition.
4. What was the enabling technological change that made photography practical for crime scene photography?
5. Digital photography was able to achieve good quality images as far back as the 1980s in space exploration and military applications. It did not begin to achieve real adoption in law enforcement until the 1990s. What are some of the factors that could have caused the delay?

6. In order for a photographic system to serve as a useful surrogate, it must be able to successfully accomplish three functions. What are those functions?
7. What are the functions of the rods and cones?
8. When we see something and say it is yellow, what can we say about the light coming from that object? What is color?
9. When we say that we see something, where is the actual image that we see?
10. The images that humans see are structured in a different way from the way that mechanical images are structured. What are the two structures?

Dynamic Range

Semiconductor light sensors are fundamental to all digital image-capture devices, including still cameras, video cameras, and scanners. Inside these devices, particles of light called *photons* will strike active sites in the semiconductor crystal and release an electron, or particle of electricity. For each electron that is knocked out of its place in the crystal structure, a *hole* is left behind. An applied electrical field will cause the electron to migrate in one direction and the hole to migrate in the opposite direction. Migration is accomplished by an electronic game of musical chairs. The loose electron displaces another bound electron, which is now free to do the same to another neighbor. The hole does the same thing in the opposite direction. The result is that an electric current flows across the crystal. When electrical charge flows, it becomes an electrical current. If current flows up to a point and collects, it causes a build-up of charge. Each electron carries a unit of electrical charge, and as more and more sites are struck, more electrons are released and more charge builds up. These devices are rated by a conversion efficiency, which is the amount of charge that either flows or builds up per unit of impinging light. So if the efficiency is 90%, then 100 units of exposure will produce 90 units of charge. Two hundred units of exposure will produce 180 units of charge, and so on. The response is linear over a range of light levels.

In situations where the level of incoming light is very low, there is virtually no build up of charge due to incoming light. However, because of thermal energy, a very small number of electrons will become free and there will be a build up of charge due simply to this occasional, accidental release. This can be seen in Figure 2.1, where a portion of a dark area has been lightened to show the random noise that results from dark current and related low-level problems.

Since the effect is thermally induced, the effect is temperature-dependent. The flow of electrons due to accidental release is called *dark current*. It is not until the flow of electrons due to incoming light is somewhat greater than the dark current, that the sensor becomes a reliable indicator of the amount of light. Below this threshold level, there is no valid indication from the device of the

FIGURE 2.1 *Colored Speckle Noise. In low exposure areas, digital photo sensors tend to exhibit random noise that is large compared to the low signal. In the figure, a dark portion of the image is brightened to show the speckle pattern inherent in that area.*

amount of incoming light. The threshold is determined by the noise level and becomes an indicator of the sensor's basic sensitivity level. Current is a measure of the flow of charge per unit time. So in a unit of time, with a single unit of current, one will accumulate a single unit of charge. Since the average dark current stays at a fixed level during a photographic exposure, and the light-induced current increases with the amount of incoming light, the signal-to-noise ratio will increase from this point on until the sensor becomes saturated.

Imagine that we have a cylindrical bucket. We pour in water at a certain rate for a given unit of time and then check the height of the water in the bucket. The height of the water is a valid indicator of the amount of water that was poured in, assuming that we stop before the bucket becomes full. Once full, all additional water will spill over the top. So, once the height of the water equals the height of the bucket, the height of water is no longer a valid indicator of the amount of water poured. The sensor chips work in much the same way. Incoming light induces the flow of electrons. The electrons

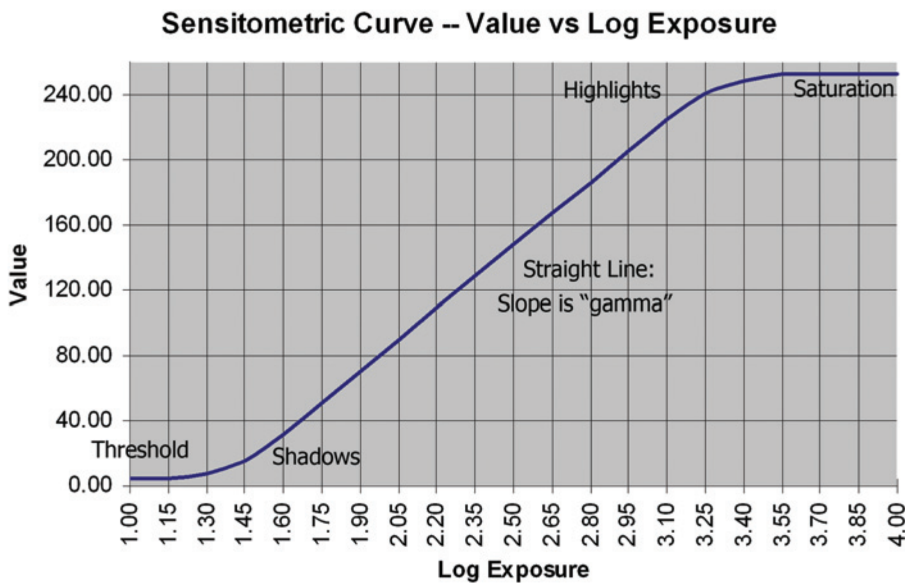


FIGURE 2.2 *Sensitometric Curve. The graph depicts an idealized response function for a digital photographic system. It shows the value output levels that result from different input light levels. The input axis is logarithmic.*

collect in small, designated portions of the sensor's surface, resulting in the collection of electrical charge. The amount of charge is a valid indicator of the amount of incoming light once above the threshold level and it remains so up to the point where the given portion of surface will not hold any additional charge. At this point the sensor is saturated and increases in incoming light will not result in an increase in electrical charge.

This explains, generally, how sensors respond. At very low light levels, below the threshold level, the sensor does not appear to respond to incoming light. From that point on, increases in incoming light result in proportional increases in the amount of charge accumulated. This type of response will continue up to the saturation point, where the sensor will hold no more charge no matter how much more light impinges. The range of light levels between the threshold point and the saturation point is the *dynamic range* of the sensor.

The sensor chips either contain devices to measure charge and produce an analogous digital number, or the charge is taken from the sensor chip and then converted to a digital number. The numbers are measures of the amount of impinging light and are called *brightness value*, or simply, *value*. Figure 2.2 shows a characterization of the response curve of a sensor chip.

There are three common ways to indicate dynamic range. For most photographers, the most common is in terms of f/stops. Lens openings are

TABLE 2.1 *Approximate Values of Scenes Under Various Conditions*

Condition	Lux	Ratio	Relative Log E	Number of f/stops
Sunlight	108,000.000000	10,000,000,000	10	33.2
Full Daylight	10,800.000000	1,000,000,000	9	29.9
Overcast Day	1,080.000000	100,000,000	8	26.6
Very Dark Day	108.000000	10,000,000	7	23.3
Twilight	10.800000	1,000,000	6	19.9
Deep Twilight	1.080000	100,000	5	16.6
Full Moon	0.108000	10,000	4	13.3
Quarter Moon	0.010800	1000	3	10.0
Starlight	0.001080	100	2	6.6
Overcast Night	0.000108	10	1	3.3
Dark Darkroom	0.000011	1	0	0.0

The Table Shows approximate lighting levels of commonly encountered conditions. The overall range is 10 orders of magnitude. The human visual system can deal with about six orders of magnitude.

traditionally measured in f/stops; in the most common series of settings, each stop represents a factor of two. That is, each successive f/stop has twice the open area than the previous one. If the dynamic range were indicated to be five f/stops, then the brightness ratio that could be accommodated would be $2 * 2 * 2 * 2 * 2 = 32$ to 1.

This brings in the next most common way to indicate dynamic range: a simple statement of the brightness ratio that can be accommodated, where the brightness levels are measured linearly. Table 2.1 shows the brightness levels for a number of common settings.

Table 2.1 indicates that full daylight brings a brightness of about 10,750 lux. At deep twilight the setting is bathed in a bit more than one lux. If a dark object were in the shade in a full daylight setting, it might reflect only about as much light as the deep twilight setting. The result is that the scene has a brightness range of at least 10,000:1, and the sensor must have a dynamic range of at least that much in order to faithfully reproduce all the elements of the scene. It is not uncommon for bright scenes to have brightness ratios of 1,000,000:1. The best commonly available sensors are color negative films specially made for portrait work. These have a dynamic range of about 20,000:1, so some compromises will have to be made.

The third way in which dynamic range might be stated is log (base 10) cycles, or factors of 10. In this terminology the ratio 10,000:1 would be stated as 4 log lux cycles.

Scanner manufacturers often refer to the dynamic range of their devices in terms of the density range to which the unit can respond monotonically. Since density is a log (base 10) unit, a density of 3.0 refers to 1/1000 of the light at density equal to zero. Due to the nature of the log scale, a density of 3.3 would be 5/10,000 and 3.6 would be 2.5/10,000. This is a legitimate dynamic range measure.

Digital camera manufacturers typically address the issue of dynamic range in terms of bit-depth. This is a related measure, but not necessarily a direct measure. If an analog-to-digital (atd) converter has the ability to distinguish 256 different levels of analog input, then it is said to have a depth of 8 (binary) bits. This is because 2 multiplied by itself 8 times equals 256. If the converter is rated at 10 bits, then the number of levels would be 1,024, or 2 multiplied by itself 10 times. The increments in output image value of the 10-bit system will be one quarter the increments of the 8-bit system. So the output scale is cut into finer increments. The bit depth is a direct measure of the fineness of the output tone scale. Imagine that the smooth curve shown in Figure 2.1 is in reality a stair-step curve. The step heights are controlled by the bit depth—the more bits, the smaller the step heights.

But, if the first step is defined by a certain signal-to-noise ratio needed to get a reliable threshold reading, and if the system is designed not to seek finer increments than the first, then the bit depth becomes an indirect indicator of dynamic range. Nonetheless, we could take a sensor that just begins to respond at 0.01 lux-seconds and saturates at 20 lux-seconds and read its output with either an 8-bit atd converter or a 10-bit converter, and the true dynamic range would still be 2000:1 (20/0.01). The important facts are that dynamic range is an indication of the input range of light that can be monotonically represented, and bit depth is a measure of the fineness of the output tone scale.

In practical terms, consider that you are taking a wedding picture. You have the bride and groom before you. The bride has spent a fortune on an elaborate dress with white lace detail superimposed on a white satin base—she is dressed in white-on-white. The groom is wearing a tuxedo with black velvet lapels on an otherwise black wool cloth—he is dressed in black-on-black. The groom's outfit requires responses to at least two very low light levels. The bride's outfit requires small differences to two very large levels of light. To add to the difficulty, the subjects are standing so that they partially face the camera and partially face each other. This makes the light coming from the lapels even lower than normal due to shadows. And, the train from the bride's gown is splayed in front of her, making that all the brighter. All of this must be in the same image. If the dynamic range of the sensor is not as great as the range of brightness in the scene, the picture will be disappointing to either the bride or the groom, or both. The wedding shot scenario is a very real problem and was partially responsible for the introduction of "portrait" films, which have extended dynamic range. If an additional light can be used to shine on the groom, it can make the result more pleasing. Such interventions are possible in studio shots, but are often difficult in the field at crime scenes.

ISSUES OF PERCEPTION OF BRIGHTNESS

Long ago, it was found that humans see equal *percent* increments in luminance as equal *perceptual* increments. Consider these situations:

1. A person is shown a gray card with a luminance of 1 unit. On that card are two smaller cards. One has a luminance of 1.5 units and the second has a luminance of 2.0 units.
2. The first smaller card is at 1.5 units and the second is at 2.25 units.

In the first case the increase in brightness is linear: 1/1.5/2.0. In each instance the absolute change is to add 0.5 units. The increments will not be perceived as equal. In the second case, the increments are proportional: 1/1.5/2.25. That is, to get from one to the next, multiply by a constant, which in this case is 1.5. These increments will be perceived as equal. The Weber-Fechner Law describes this phenomenon and holds that equal ratios of luminance increases are perceived as equal increments. That is, over a wide range of luminance levels:

$$\text{(Increase in Luminance)} \div \text{(Base Level of Luminance)} \\ = \text{Constant Perception or Increase}$$

For well-designed viewing conditions, the constant is about 0.01. That is, 1% increases will be seen as equally brighter. Increments greater than 1% will also appear perceptually the same. For example, if the increments were both 50%, those increments would be seen as the same. Note that in comparing dark and bright settings the differences in absolute change are quite dramatic. For example, if dark areas are at about 1 lux, an increase of 0.01 lux would be seen as an equal increment compared to a bright section of 1000 lux, where the increment is 10 lux, ten times the level of the dark area base.

To expand the basic relationship to cover a full spectrum of situations, we integrate the point relationship over the full range of perceptions, P:

$$P = \int dL/L = \ln(L)$$

That is, the perception of brightness is related to the logarithm of absolute luminance. Normally, in photography the base 10 logarithm (log) is used instead of the natural logarithm (ln), and the relationship is the simple:

$$\log(L) = 2.3 * \ln(L)$$

This all goes to show that in photography, where a mechanical set of devices serves as a surrogate for human viewing at the scene, it is appropriate to represent the brightness' of portions on an image on a logarithmic scale. Similarly, most of the settings on photographic devices work in equal ratio increments, usually factors of two.

BEER'S LAW

In keeping with the spirit of the founder's name, assume that we are in a bar and that this bar serves beer in rectangular glasses. Looking down at the glasses from the top we see that they have a width, W, and a thickness, K. All the beer served in this bar is well filtered so that light going though the glasses of beer does not scatter. The beer is essentially a solution of some special, colored materials in water. Some of those materials have molecules that absorb light in the blue and green portions of the visual spectrum. So the beer has an amber color. Since the liquid has a certain number of molecules of absorbing material distributed evenly and randomly throughout, and since the light is composed of a stream of particles of light called *photons*, the probability that a photon will hit and be absorbed by a molecule of colorant (or dye) is proportional to the number of such molecules per unit volume. That is, the amount of absorption depends on the number of such molecules per liter of liquid: the more molecules per liter, C, the more absorption, A. This is a straight linear relationship:

$$A = a \star C$$

The "a" is a constant that is dependent on units of measure, spectrum of light, and nature of the chemistry involved. Absorption is represented as a percent of the light that is coming into the system. The inverse of A is Transmittance, or T, where:

$$T = 1/A, \text{ conversely, } A = 1/T$$

Continuing the experiment, if we make a set of measurements with beer right out of the tap, we would find a certain level of absorption per glass. If we let the beer sit in a pitcher on the bar till one-half of the base liquid had evaporated and then repeated the measurement we would find twice the absorption. This is because the water evaporated away and the colored material was left behind. Since half the water is gone and all the colorant is still present, the concentration of the absorbing material has doubled.

Now assume that the peculiar, rectangular beer glasses are very thin. That is, the thickness, K, is small compared to the width W. We already have found that for a single glass of beer, a certain percent of the light that impinges on the glass is transmitted and comes out the other side. If two glasses were set next to each other so that the light coming through the first glass was then set to go through the second, the resulting transmittance would be the product of the two separate transmittances. That is if 30% of the light came through each glass taken alone, then the combined effect would be 30% of 30%, which results in 9% (0.30 * 0.30 = 0.09). If there were three glasses, the result would be 2.7% (0.30 * 0.30 * 0.30 = 0.0 27). If there were one special glass that had three times the thickness of the normal glass, it would be equivalent to three normal glasses in series (there are some special factors that will be considered in later chapters). This indicates that the absorption is proportional to the thickness of the absorber, K. Combining this with the earlier finding with respect to concentration:

$$A = a * C * K$$

This is the basis of Beer's law. The absorption of a nonscattering material is proportional to its concentration times its thickness. Since photographic systems are best represented in logarithmic terms, this can be restated as:

$$\text{Log}(A) = \text{Log}(a * C * K), \text{ or}$$

$$\text{Log}(T) = \text{Log}(1/a * C * K) = -\text{Log}(a * C * K)$$

DENSITY

The most common way to measure photographic prints is in terms of density, D, which is defined as the Log(1/T), or – Log(T). Apply this to Beer's law:

$$D = -\text{Log}(T) = \text{Log}(a * C * K)$$

That is, the density of a patch is proportional to the concentration of dye in the patch times the thickness of the patch.

Remembering that:

$$\text{Log}(X * Y) = \text{Log}(X) + \text{Log}(Y)$$

if two transmissive patches are used in series, the resulting density will be the sum of the densities of the two patches. In traditional silver halide photography, the concentration of dye is controlled by the creation of dye molecules during film processing and in response to initial film exposure in the camera.

SENSITOMETRY

There are standard ways to convey the response of photographic systems. These were initially developed for use with film photography and then adapted for digital photography. The basic response of a photographic system often is represented in a graphical form as shown earlier in Figure 2.1. This approach was developed over 100 years ago by Ferdinand Hurter and Vero Charles Driffield. The curve is called the Hurter-Driffield (or H&D) curve, or more descriptively, the sensitometric curve, response curve, D-Log E, or characteristic curve. In the cases of both film and digital photography, the vertical axis is indicative of the response of the sensor or sensor system, and the horizontal axis is the amount of light impinging on the sensor. So it shows the transfer function or the amount of output for each level of input.

In the case of film photography the output is density, which is equal to the logarithm of one divided by the fraction of light reflected (reflectance) of a reflection print or the transmittance of a transparency. It gives higher numbers for lower levels of light from the sample. It is also a logarithmic scale, and so is consistent with how people see and is capable of easily showing a very wide range of values. In the typical film photography system, more impinging light results in higher densities so that increases along the vertical axis indicate darker patches on the output. (Photographic slide films are plotted on the same axes as negative films, but the sample patches get lighter with increasing impinging light.) In representing film systems, the convention is to show the input, or impinging light axis as the logarithm of exposure. Again it is convenient to do this because the system can cover a very wide dynamic range and because the human visual system responds logarithmically. Film sensitometry is shown as a log-log plot: Density vs Log E, hence the descriptive jargon, D Log E.

The convention for digital photography is somewhat different. First of all, unless otherwise indicated, the vertical axis is linear. The output or vertical axis is "value," which as indicated earlier, is linearly proportional to the response of the sensor (the amount of charge accumulated). The horizontal axis is usually Log E, as with D-Log E curves. Since the human visual system is logarithmic, digital image output values need to be converted to log values to be seen as normal by humans. If this had already been done, then graphical representations shown in image-editing software might simply show images that are the inputs to the editing process and the corresponding output values, already in logarithmic terms. One needs to be careful in interpreting these graphs. Note that film sensitometry has a vertical scale where increases in the output axis indicate increasing image darkness, whereas in digital sensitometry, increases along the vertical axis are increases in image brightness. Digital camera response is somewhat similar to that of slide film, but plotted upside down.

There are a number of portions of the characteristic curve that have special significance and are named as indicated in Figure 2.1. The *toe* refers to the areas that are bright, but not on the flattened portion of the curve that extends beyond the saturation point. The *shoulder* refers to the portions that are dark, but not on the extension beyond the threshold point. Note that with negative film, the toe will be in the lower left of the graph and the shoulder will be in the upper right. The reverse is true for digital cameras. With slide films, the toe is in the lower right and the shoulder is in the upper left.

Film typically has a graceful transition from the sloped portion of the curve to the flat portions, whereas digital cameras typically have a sharp toe transition and a graceful shoulder. This is because as the sensor fills with charge, it suddenly reaches the point where it can hold no more, and there is a sharp cut-off. In the shoulder, as the current due to light approaches that due to accidental dark current, there is a more asymptotic behavior. The details in the bride's gown are rendered in the toe, and the details in the tuxedo are rendered in the shoulder.

These portions of the curve also are referred to as the highlight and shadow portions, respectively. With a sharp toe, it is important that the photographer not overexpose the photo since detail will quickly vanish. Some recommend that the photographer purposely seek a slight underexposure setting. However this could jeopardize the shoulder. Instead the photographer must be really careful to get the optimal exposure for the scene or run an exposure series. Slide films are similarly sensitive. Negative films, especially the so-called portrait films, are significantly more forgiving. This ability to be forgiving is sometimes referred to as *latitude*.

In color images, there are separate records, one each for red, green, and blue portions of the spectrum. Each will have its own characteristic curve. If the toe of the image has gradual red and blue records and a sharp green toe, highlight portions of the image will have a magenta cast (magenta is the lack of green). This effect, when referring to Caucasians, is called *beefy flesh tones*. If the green and blue records are gradual and the red is sharp, the flesh tones will be cyan (cyan is the lack of red), or *cadaverous*. Comparable effects can occur in the shoulder but they are less noticeable. It is generally desirable for the three records to have the same curve shape. This makes it much easier to adjust the color balance in an image.

GENERAL CHARACTERISTIC CURVE DESCRIPTORS

Brightness and contrast are the most common descriptors that people use when describing photos. One (digital) image will appear to be brighter than another if the characteristic curve is shifted upward.

Figure 2.3 shows two characteristic curves plotted on the same graph. Both have the same shape and horizontal positions, but one is higher than the other. It will appear brighter. If one of the color channels is shifted upward relative to the others, the image will have an overall color cast. So if the red curve were higher than the green and blue curves, the overall picture would have a reddish cast.

Figure 2.4 shows two characteristic curves. One has a steeper slope than the other in the central portion of the curve. That image will appear to have more contrast than the other. Dark-to-light ratios will be exaggerated. Dark areas will be darker and light areas will be lighter, with the result that differences in brightness for different parts of the image will be enhanced in high-contrast versions. In the extreme, fully increasing the contrast will result in an image that has only black and white, with no intervening shades of gray. Reducing the contrast to zero will result in an image with no content—everything is a middling gray. If one of the color curves is shifted relative to

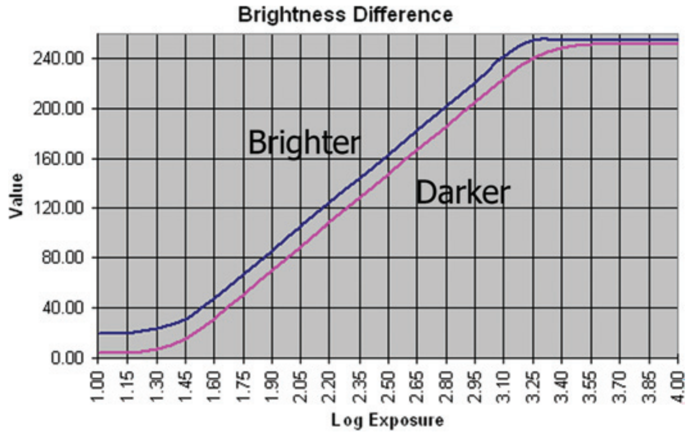


FIGURE 2.3 *Brightness Difference. The two graphs show a darker image and a brighter image. The shift is strictly a vertical displacement.*

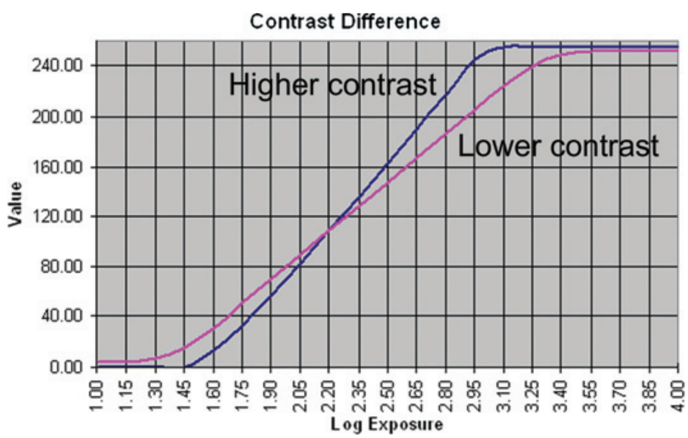


FIGURE 2.4 *Contrast Difference. The two graphs show a response with more output per unit input—high contrast, and less output per unit input—low contrast. The shift is a slope displacement.*

the others, the image will have a color mismatch that varies with overall scene brightness. For example, if the green curve were to be shifted to lower contrast, the toe would be greenish and the shoulder would gain a general magenta cast.

The slope of the mid-scale or "straight-line" portion of the curve is referred to as *gamma*. When gamma equals one on a log-log plot, such as a D-Log E curve, there is a one-to-one relationship between input brightness and output brightness. At all other values of gamma, the relationship is nonlinear. In the 1980s the Eastman Kodak Company conducted a study of consumer preferences and found that even though the engineers preferred the logic of the gamma = 1 approach, consumers preferred prints in which the gamma was a bit greater than 1. They promptly changed the gamma of their amateur negative films.

In a multiple-stage process, such as film photography where the camera creates a negative that is then printed onto a sheet of photographic paper with a similar characteristic curve, the gamma of the overall system is the product of the gammas of the individual parts. To make the original camera work more forgiving in the field, it is common to make the camera sensor gamma relatively low. This makes it easier to get a greater dynamic range, and makes it possible for the photographer to allow for some error in setting the exposure level without ruining the photo. To compensate, the subsequent processing of the image requires a higher gamma for the print material. This will make the overall system gamma is a bit higher than 1. For example, the gamma of a motion picture negative (camera) film could equal 0.5, and the gamma of the print film 2.2. The result will be an overall system gamma of $0.5 * 2.2 = 1.1$.

The threshold point was described earlier as a level on the input axis that reliably results in a response from the sensor that is not primarily random noise. Increases in input light beyond that point result in monotonically higher output responses. This point can serve as a speed point; that is, a single number that is a measure of the sensitivity of the sensor. When the level of exposure needed to get to the threshold point is, for example, 5 lux-seconds, then the speed of the sensor would be some multiple of the inverse of this number, or 0.2. The inverse is used in order to make higher numbers indicate higher sensitivity. A nice property of this convention is that when the speed point times the exposure is equal to 1.0, the sensor will be properly exposed. Most commonly, the logarithm of exposure is used instead of the absolute number since photographic systems really need to be logarithmic. The logarithm of 1.0, of course, is zero. So when considered in log space, when the speed point (in log space) plus the exposure level (in log space) is equal to zero, the camera is set for proper exposure.

Photographic speed points normally are stated according to a formula set by the International Standards Organization, and therefore are referred to as ISO values. On the log scale, equal multiples of the speed ISO will result in equal multiples in the amount of sensitivity. Accordingly, a setting of ISO 200 will be twice as sensitive as one of ISO 100. Likewise a setting of 400 will be twice as sensitive as one of 200, and four times as sensitive as one of 100. This translates into other settings as well. If a digital camera is set to ISO 400 and it gives proper exposure at 1/100 of a second, the camera could be reset to ISO 100 and 1/25 of a second and give the same level of exposure.

Figure 2.5 shows two characteristic curves that have the same shape, but one is shifted horizontally relative to the other. The sensor depicted by the curve on the left is more sensitive than the one on the right; in other words, it starts to respond reliably to light at lower levels. If the curve on the right

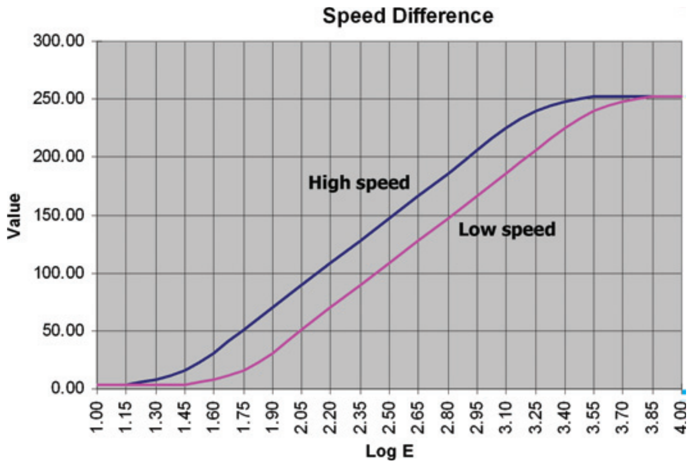


FIGURE 2.5 *Speed Difference. The two graphs show two systems that vary in sensitivity or speed. The shift is strictly a horizontal displacement.*

responds at 34 lux-seconds (incident on the subject) and the one on the left at 17 lux-seconds, then the difference in the ISO ratings for the two sensors will be different by a factor of two as well. For example if the one on the left is ISO 200 then the one on the right is ISO 100. If a photographer were taking photos under the two conditions and the same lens opening was chosen, then the exposure times might be 1/100 of a second for the lower sensitivity sensor and 1/200 of a second for the other. The relationship between sensitivity and shutter opening time is what gives rise to the term *speed*. The higher ISO film is *faster* than the other.

To summarize, if two characteristic curves are displaced vertically relative to each other, the higher one will produce a brighter image. If one has a steeper slope than the other, it will have higher contrast. And, if two curves are displaced horizontally, the one on the left will have more sensitivity and a higher ISO.

One important point to notice. If image information falls on the sloped portion of the sensitometric curve, it can be made to show in the final image. If it falls on the flat portions of the curve, before the threshold point or after the saturation point, the information will not be recorded. With digital image processing a lot can be done to bring out weak information, but there is no way to enhance information that is not recorded in the first place. Going back to the wedding picture, if the white-on-white weave in the bride's dress is beyond the saturation point, it cannot be rendered. And, if the black velvet lapels on the groom's tux are below the threshold level, they cannot be rendered. No amount of image processing will help. The lapels must be above the threshold level, and at the same time, the dress must be below the saturation point if both are to show in the same photo. And this must hold for all three primary color channels.

EXERCISES

The following exercises involve using Adobe PhotoShop software tools associated with some of the topics discussed in this chapter. All images not included in the text can be downloaded from the web sites.

Levels Dialog Box

Open the **image** > **Image Size.jpg**. Select **Image** > **Adjust** > **Levels**. The Levels dialog box will appear. Click the Preview checkbox. In order to use the Levels control accurately, you must understand what each control is doing to the image.

All tones in an image are represented by a histogram in the Levels dialog box as shown in Figure 2.E1. A total of 256 tones are shown on a horizontal scale of 0 to 255. 0 represents a full black in the image and 255 represents a full white.

The height of each bar indicates the number of pixels at that brightness level.

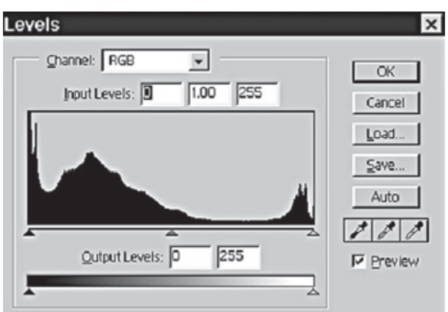


FIGURE 2.E1

Compare the histogram to the image in Figure 2.E2. The histogram represents the entire image. The dark areas (shadows) are represented on the left and the light areas (highlights) on the right. Notice in this image, there are many more dark pixels than there are light pixels as represented by the histogram.

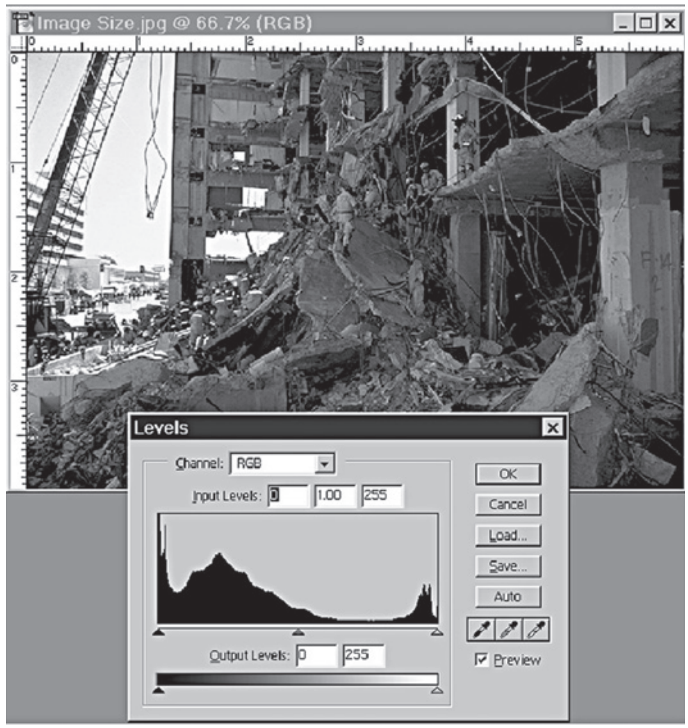


FIGURE 2.E2

The dialog box is divided into two sections:

- Input levels, which are used to adjust contrast in the midtones (gamma).
- Output levels, which generally are used to reduce contrast in the highlights and shadows of the image.

Always adjust the input levels first; this will adjust the contrast (gamma) and the density. Moving the black point and white point sliders on either end inward spreads the range of brightness, increasing the contrast.

Drag the black slider to the right—the dark areas get darker, thereby increasing the contrast (see Figure 2.E3). Return it to its original position. Drag the white slider to the left—the white areas get lighter, thereby increasing the contrast (see Figure 2.E3). Return it to its original position.

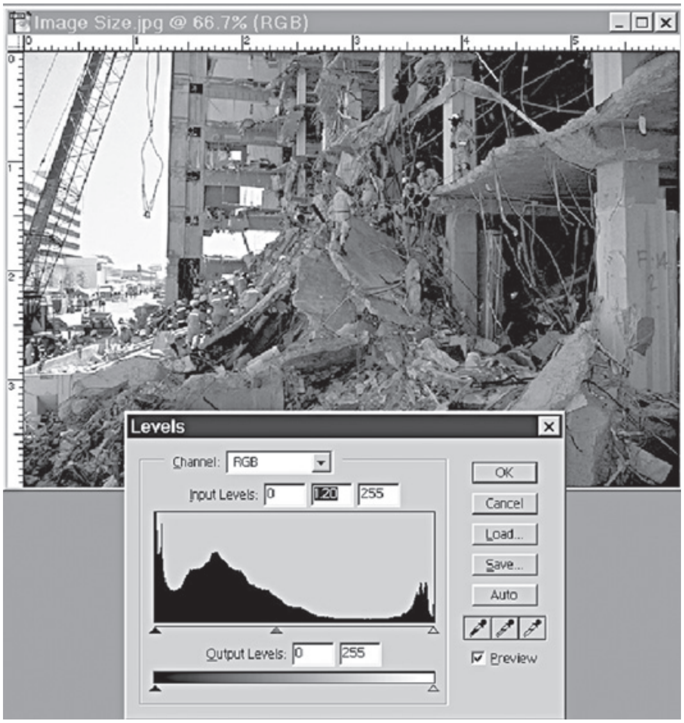


FIGURE 2.E3

This image has some lost detail in the blacks, but the white is acceptable. The detail lost in the shadow areas cannot be recovered. Overall, the midtones in this image are too dark. The gamma control in the center of the input levels allows you to adjust the brightness of the midtones without affecting the shadows or highlight areas.

Move the Gamma slider to the left until the input indicates 1.2. Notice the midtones were lightened, showing more detail in the darker portion of the image.

In case you make a mistake, all settings can be returned to their default position by holding the **ALT** key, which changes the **Cancel** button to a **Reset** button.

The output levels are always adjusted last. These adjustments are used to reduce contrast in the highlights and shadows of the image (see Figure 2.E4). The highlights (light areas) of this image have pretty good detail, but some detail is lost in the shadows. Slide the black slider to the right, increasing the detail in the shadows until 20 is indicated in the window. This lightens the shadow areas (decreasing the contrast).

Note: If you completely switch positions of the left and right sliders, the image becomes a negative.

Auto Levels

PhotoShop also has an Auto Levels control available through the **Image** > **Adjust** menu or the Levels dialog box. Always try the Auto setting on the image—it can save you some time. If it doesn't correct the image, reset it and do it manually.

In this case, very little is changed when the **Auto** button is pressed, basically because most of the pixels in the image were dark anyway.

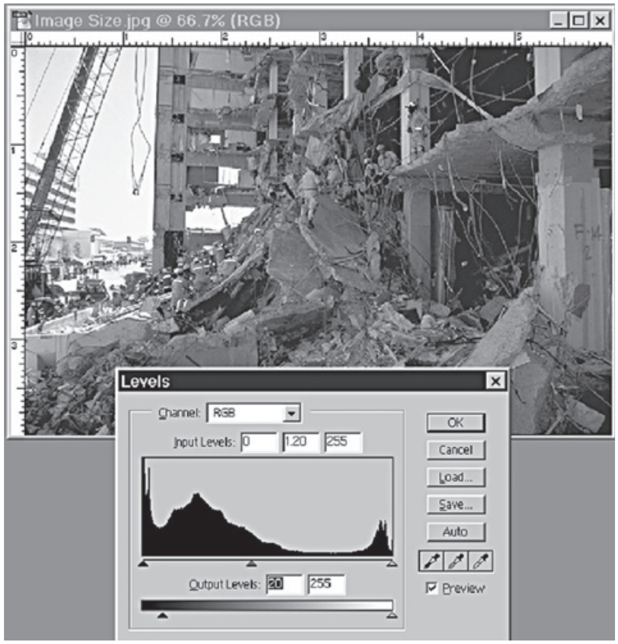


FIGURE 2.E4

For this to work effectively you need a full black–full white and a good range of midtones indicated in the histogram. This image has many more black pixels than it does white and so does not work well using the **Auto** adjustment.

Using the Eyedroppers

In addition to the manual and auto adjustments in the levels control, the image can also be adjusted using the Eyedroppers to sample black, white, and a midtone gray (18% gray is the standard) in the image. If used properly this technique will not only adjust the tones but also the color balance of the image.

A problem arises with this technique if there is no black, white, or gray in the image.

Open the two images, DCP 00164 and DCP 00165, and place them side-by-side as shown in Figure 2.E5. Look closely at the two images. The tank tops are two different colors, although it is the same individual photographed basically at the same time.

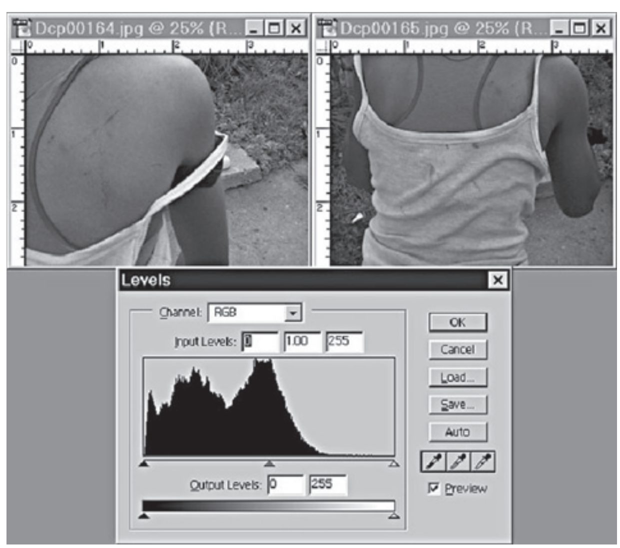


FIGURE 2.E5

Locate a black, white, and gray in the image:

- There is a black strap around the right arm.
- The strap of the shirt is white.
- A good substitute for 18% gray is concrete.

Select the black eyedropper, move it around the black strap to find a very low and relatively similar value for the red, green, and blue values, and click on the black strap. To see the levels, be sure the Info palette is open. If it is not, go to the **Window** menu and select **Info.**

Select the white eyedropper, move it around the white strap to find a set of three similar numbers close to 255. Click on the white strap.

Select the gray eyedropper, move it around the concrete to find a set of similar numbers that are midtones. Click on the darker portion of the concrete.

The colors have been balanced; midtones may still need adjustment by using the gamma slider. Repeat this same procedure for the remaining image. The colors of the shirts should match closely when you are finished.

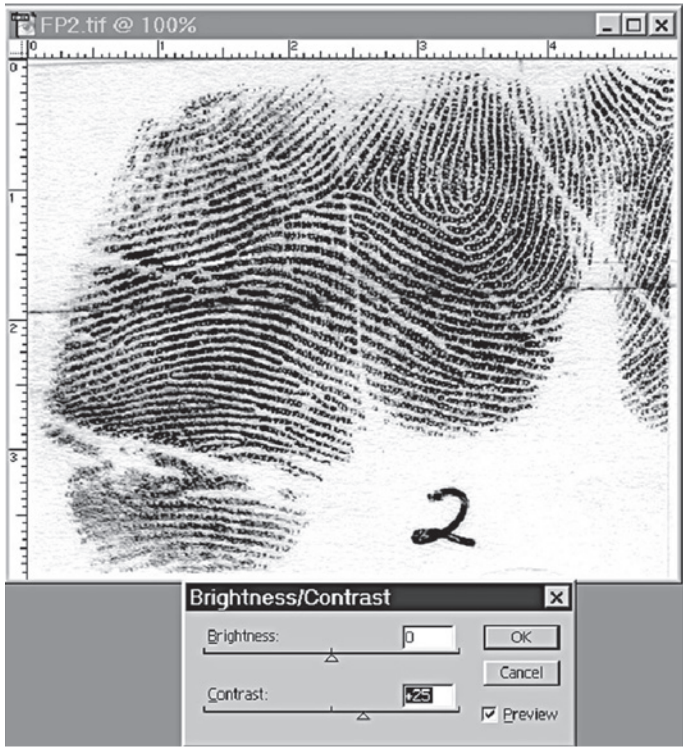


FIGURE 2.E6

Brightness and Contrast

Using the Brightness/Contrast command is the easiest way to make simple adjustments to the tonal range of an image, but it offers very little control over setting exact highlights and shadows, while controlling the midtones of an image. In addition, the Brightness/Contrast command does not work with individual color channels and is not recommended for high-end output. Therefore, it should not be used on photographic images (continuous tone images). PhotoShop has much more powerful tools for image control.

It does, however, work quite well on images that have few midtones, such as some fingerprints and documents. Download the fingerprint image, Print_2. Select **Image** > **Adjust** > **Contrast/Brightness** (see Figure 2.E6).

Values of the brightness and contrast controls range from +100 to –100. Moving the adjustment to the left decreases the level and to the right increases it. The number at the right of each slider value displays the brightness or contrast value. You can see the changes, and you can see that some of the highlight and shadow information disappears when significant adjustments are made in these controls.

Curves Tool

The Curves Dialog Box

Like the Levels dialog box, the Curves dialog box lets you adjust the entire tonal range of an image. But instead of making adjustments using only three variables (highlights, shadows, and midtones), with

Curves you can adjust any point along a 0-to-255 scale while keeping up to 15 other values constant.

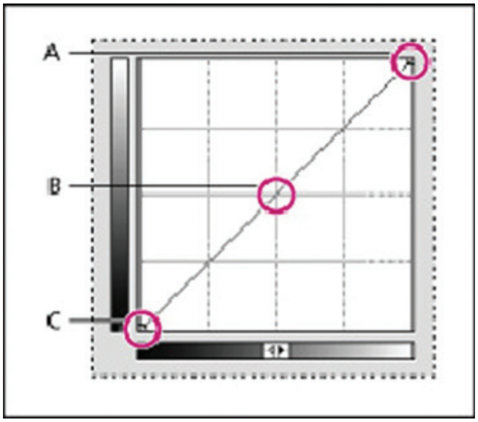


FIGURE 2.E7

You can also use Curves to make precise adjustments to individual color channels in an image as shown in Figure 2.E7. The vertical axis represents the new color values (output levels), and the horizontal axis shows the original intensity of the pixel values (input levels).

All pixels have identical input and output values shown by the default diagonal line.

1. Click any point(s) on the curve that you want to remain fixed. For example, if you want to adjust the midtones while minimizing the effect on the highlights and shadows, click the quarter and three-quarter points on the curve.
2. You can add up to 16 control points to the curve, locking those values. To remove a control point, drag it off the graph or select it and press **Delete**. You cannot delete the endpoints of the curve. See Figure 2.E8.
3. Open Image_01.jpg. Zoom in on the image so you can see the ridges and valleys clearly.
4. To determine the lightest and darkest areas in the image, drag the cursor over the image. The intensity values of the area under the pointer, along with the corresponding location on the curve, are displayed in the Curves dialog box.

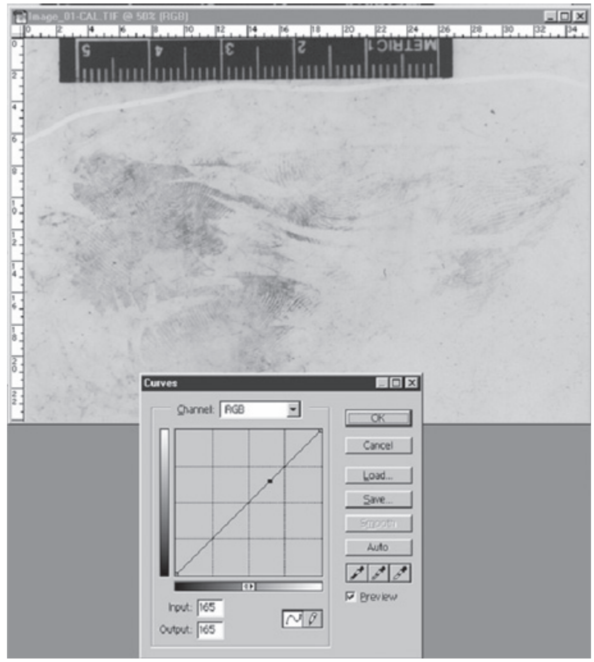


FIGURE 2.E8

5. Move the cursor (eyedropper) over the light area (valleys) of the image. Click to sample the image and notice the value in the input box.
6. Click anywhere on the curve to establish a plot point. Enter the reading from the eyedropper in the input value (approximately 192).
7. Move the cursor over a ridge (dark area) of the fingerprint. Click to sample the tonal value, and note the reading.
8. Click anywhere on the curve to establish a plot point. Enter the reading from the eyedropper in the input value (approximately 90).
9. Select the upper plot point on the curve and enter 200 as the output value. Select the lower plot point on the curve and enter 70 as the output value. Click OK to apply the changes.

This technique causes an increase in contrast by adjusting the exact tones of the image without affecting the rest of the image. That is, unlike the brightness/contrast tools, no data is lost!

FIGURE 2.E9

The image may be adjusted in many ways. Do any of the following to adjust the curve:

- Drag the curve until the image looks the way you want it.
- Click a point on the curve, and enter input and output values for the point.
- Select the pencil at the bottom of the dialog box, and drag to draw a new curve. You can hold down **Shift** to constrain the curve to a straight line, and click to define endpoints.
- When you're finished, click **Smooth** if you want to smooth the curve.

HDR (High Dynamic Range) Tool

The dynamic range is the ratio between dark and bright regions in the visible world. It far exceeds the range of human vision and of images that are captured by cameras and printed or displayed mechanically. Human eyes can adapt to many different brightness levels at virtually the same time, but cameras cannot, and computer monitors can capture and reproduce only a fixed dynamic range. Photographers working with digital images must be selective about what's important in a scene because they are working with a system with a limited dynamic range.

The High Dynamic Range (HDR) tool opens up a world of possibilities because it can represent the entire dynamic range of the visible world. Because all the luminance values in a real-world scene are represented proportionately and stored in an HDR image, adjusting the exposure of an HDR image is like adjusting the exposure when photographing a scene in the real world.

Using the HDR Tool

To make the photographs:

- Mount the camera on a tripod for direct registration between pictures.
- Manually focus the camera.
- Set the camera to manual mode.
- Make at least five photographs at varying exposures.
- Adjust the exposure with the shutter speed. Do not change the f/stop since this will change the depth-of-field (i.e., focus at various points in the image).

The **HDR** example was photographed at F/4 with shutter speeds at 1/16″, 1/8″, 1/4″, 1/2″, and 1″ sec. The photos range from underexposed to overexposed (bright to dark), as shown in Figure 2.E9. The images are in the folder HDR Images on the website.

1. In PhotoShop, select **File** > **Automate** > **Merge to HDR**.
2. Select either **Files** or **Folder** from the dropdown menu. Click **Browse** to select your **Files** or **Folder** (Figure 2.E10). All images in the folder you have selected for the HDR merge will show up.
3. Select **OK** in Figure 2.E11 to import the images.
4. The images that you want to merge can be selected by using the checkboxes on the left in Figure 2.E12. Select all the images for merging.

The merged image appears in the dialog box along with a histogram of the image as shown in Figure 2.E12.

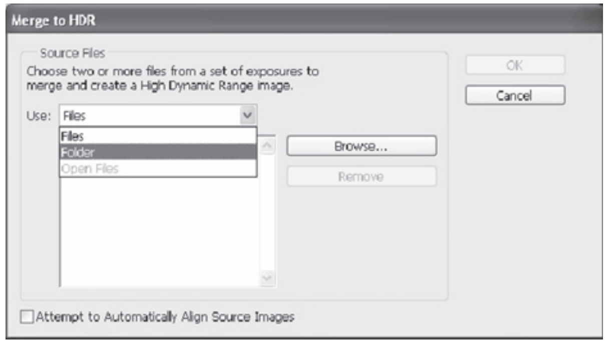


FIGURE 2.E10

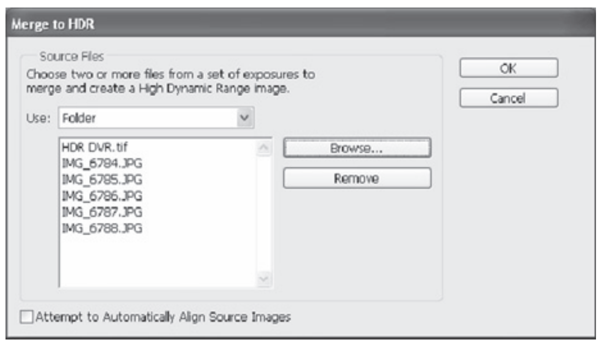


FIGURE 2.E11

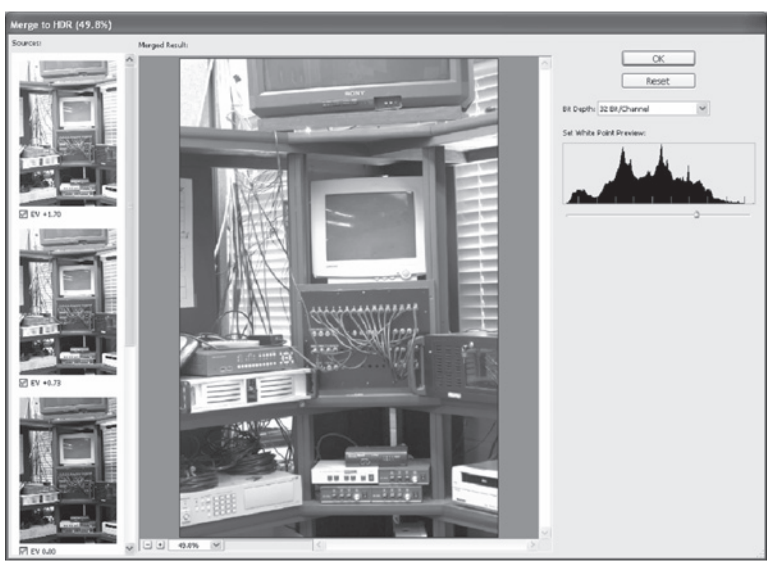


FIGURE 2.E12

Note: It is important to set the White Point of the image after it is merged. Setting the white point of the merged image allows you to adjust the mid-tones of the image before saving. The white point is set with the slider under the histogram as shown in Figure 2.E13.

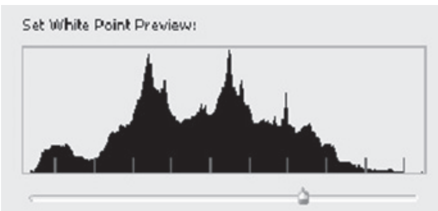


FIGURE 2.E13

Now you are ready to bring the image into PhotoShop. Note the bit depth above the histogram (see Figure 2.E14).

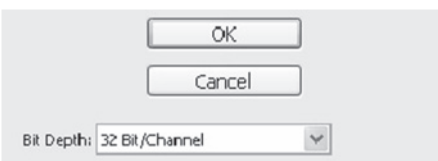


FIGURE 2.E14

Most camera images will be brought into PhotoShop at 32 bits. If you select the drop-down menu, you also have the options of 8 bits and 16 bits. This creates a problem, because if you intend on doing more adjustments to the image in PhotoShop, there are some tools that will not work with 16- and 32-bit images. Also if you intend to archive the image in the .JPG format, it recognizes only 8-bit images.

It is recommended that the image be opened in 32-bit format to accomplish the adjustments you can with the active tools in PhotoShop. (Note: The file size will be very large, in excess of 70MB.) Then if necessary the image can be converted to 8 or 16 bits inside PhotoShop.

Select **OK** as shown in Figure 2.E14.

Converting the Image to 16 Bits

Select **Image** > **Mode** > **16 Bits/Channel** as shown in Figure 2.E15. The HDR conversion window will open as shown in Figure 2.E16.

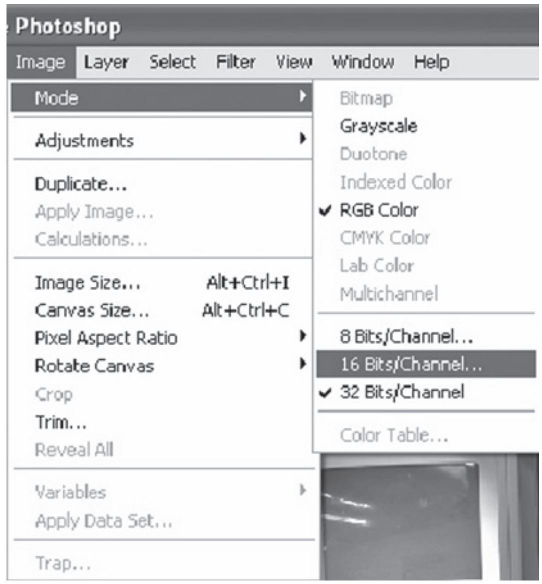


FIGURE 2.E15

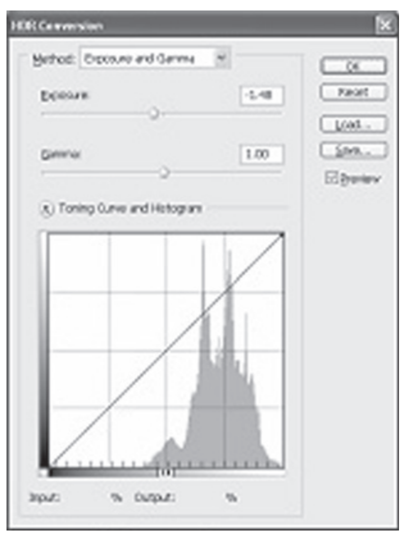


FIGURE 2.E16

Note the controls under the Method dropdown:

- Exposure and gamma
- Highlight compression
- Equalize histogram
- Local adaptation

The only adjustments that are really necessary are exposure and gamma. This can be done after the image is converted but then you would be working on a 16-bit image.

The adjustments will be more accurate if they are a 32-bit image prior to conversion. After the adjustments are made, select **OK**. You now have a much smaller image size and the majority of the tools in PhotoShop are available. The only options you have for saving a 16-bit image is .TIFF or .PSD. To save the image as a .JPG, it must be converted to an 8-bit image.

Light and Lenses

BASICS OF LIGHT

The English word *photography* is most commonly understood to mean, "writing with light," as derived from the Greek, Φως, "phos," the root of the word for light; and Γραφω, "grapho," the base of the verb, to write. So it is logical that to better understand the processes involved, the nature of light should be examined, as well as the device that allows it to enter the camera. After all, it is the light that is doing the writing.

A good start to learning about light is to consider how it is generated and what it is made of. The place to start is the simple Bohr model of the atom, which is made up of a nucleus surrounded by a group of moving electrons. In this simple model, the electrons are in circular orbits around the nucleus similar to the way in which the planets circle the sun (see Figure 3.1). Newer models of the atom are more complex, but the basics of light generation are similar enough so that we can use the simple model. In the atom, the orbits have distinct energy levels. If energy is applied to an atom at the right level, it can cause an electron to move from a near orbit to a more distant orbit. But this is not necessarily a stable condition, and the electron will eventually fall from the high energy orbit to one at lower energy, often the one from which it came in the first place. As the atom goes from a higher energy state to a lower one, it must give up some energy. This is done by emitting light. Each such transition produces a packet of energy in the form of an electromagnetic wave packet called a photon. The photon is considered a particle of light. When a charged particle (in this case, an electron) moves rapidly, it results in the creation of an electromagnetic wave. This is what is happening in the atom. Electromagnetic waves do not need a medium in which to propagate (as water waves and sound waves do). The photon will travel in a straight line until it is either sent off course by a physical obstacle or absorbed. Photons are absorbed by the reverse of the process by which they are created. They impart their energy to an electron and send that electron to a higher energy orbit. The speed of the propagation (in a vacuum) is a

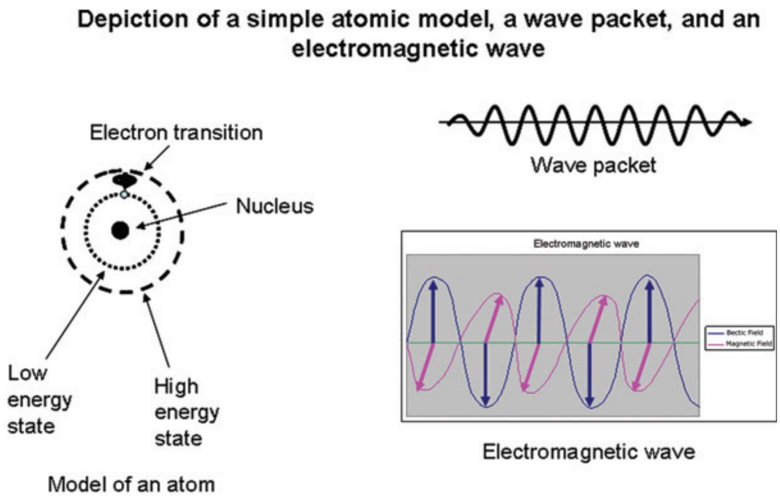


FIGURE 3.1 *Depiction of the Generation and Nature of Light. On the left is a simple model of an atom showing that the act of an electron transitioning between orbitals/energy levels causes the creation of an electromagnetic wave. Shown at the bottom is an electric wave (vertical) and its corresponding magnetic wave (horizontal) pushing each other along. Above this is a shortened wave packet, showing that the photons do not have infinite length but have a beginning and an end.*

universal constant, the speed of light, c. In materials other than a vacuum, the speed of light will be lower, and the ratio of c divided by the speed in the particular medium is called the *refractive index*. This is a property of the specific material.

A light source is a device that causes the release of photons. In an ordinary light bulb, electricity causes the metal filament to become very hot so that the atoms in the metal are violently banging into each other at a rapid rate. This causes the atoms to take on energy and move electrons to higher energy orbits. After a brief time those electrons will fall back to their ground state and emit a photon. The process repeats as long as power is applied. The emission of photons is a random process in which millions of atoms are active and the amount of energy is different from atom to atom. The result is a flotilla of photons emitted per second, and although each photon has its own level of energy, in aggregate they cover a wide range of energy levels. The energy of a photon depends on its wavelength. The property that humans call color is attributable to wavelength. Put this all together and we can say that the light from the light bulb will cover a wide range of wavelengths and colors. There are obviously other mechanisms that can result in the creation of light, such as fires, the atomic processes of the sun, the activation of phosphors in fluorescent tubes or computer monitors, and so on. The basic story is the same, but the processes differ a bit as do the colors produced.

If we were to increase the level of electrical power to the light bulb, it would glow more brightly and produce more photons per second per unit area. Also, the increase in power would cause the filament to get hotter. This means that the atoms are banging into each other with more energy on average. The higher energy in some of the electron transitions means that those cases will impart greater energy to the resulting photons. Higher energy in photons is accompanied by shorter wavelengths. Humans perceive the higher energy, short-wavelength photons as blue. A bit lower and they become green, then yellow, and finally red. So if we put more electrical power to a light bulb, the light becomes brighter and more bluish. Conversely if we decrease the power to the bulb, the light will become less intense and more reddish. This effect is conceptualized as *color temperature*. It turns out that the higher the temperature of a black body source, such as the filament, the more blue light there will be compared to red light (see Figure 3.2 for approximations of light outputs from typical sources). So, for example, we can say that the color temperature of normal, midday, northern daylight is about 6000 degrees Kelvin (°K) and a normal incandescent light bulb is about 3000°K. Just knowing the one number gives the spectral quality of the light from the source.

The concept of color temperature is derived from the work of Max Planck, who advanced the theory of light emission and derived the basic

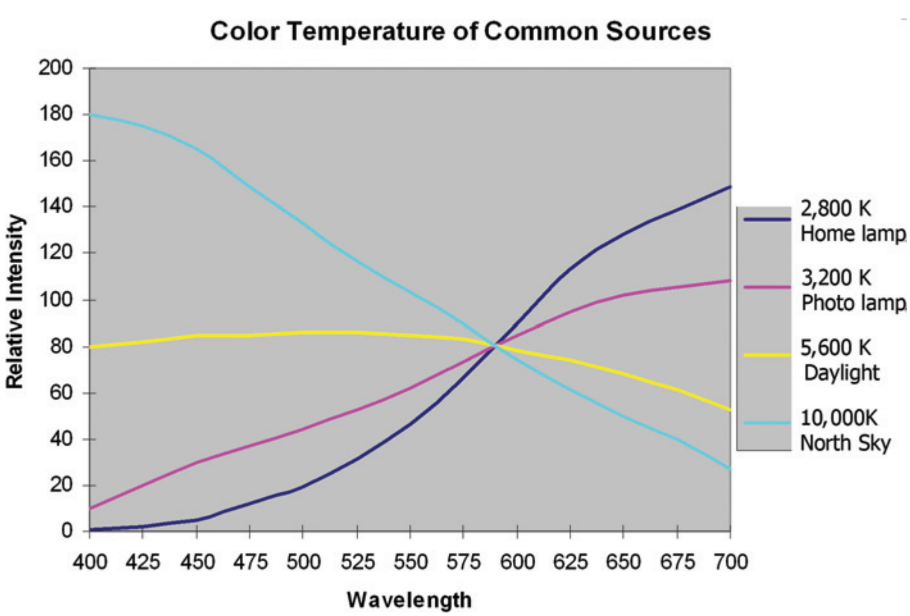


FIGURE 3.2 *Color Temperatures of Common Sources. The curves show the approximate energy distribution by wavelength for each of several common light sources.*

equation for determining the amount of energy available at each wave length from a black body radiator or a certain temperature.

$$S(\lambda) = \left[\frac{8\Pi hc}{\lambda^5}\right]\left[\frac{1}{e^{(hc/\lambda KT)} - 1}\right]$$

where:

- h is Planck's constant, 6.626 exp -34 joule seconds
- c is the velocity of light, 2.99792 exp 14 microns per second
- K is Boltzman's constant, 1.380622 exp -23 joules/degree Kelvin
- λ is the wavelength in microns
- T is the temperature in degrees Kelvin
- Planck's equation gives energy values in S_λ watts/square meter at each wavelength, λ

The values for the numerical constants are dependent on the system of units being used. Planck's equation is a cornerstone of physics, but use caution when employing it. Most light-measuring devices, such as cameras and light meters, actually are photon counters. They do not measure energy, that is watts per square meter, but rather they count photons per second per square meter. And since they have a fixed area of exposure and read for a fixed time, they simply report the photon count for that area and exposure time. To convert Planck's equation to photons per second per square meter, simply divide the basic equation by the energy of each photon, which is given by:

$$\text{Energy per photon} = h * c/\lambda$$

This gives the energy of a photon in joules. And since a watt is equal to one joule per second, the modified equation gives the number of photons per second per square meter. The result is:

$$M(\lambda) = \left[\frac{25.1327}{\lambda^4}\right]\left[\frac{1}{e^{(14{,}383/\lambda T)} - 1}\right]$$

In this statement of the modified Planck equation, the constants already have been evaluated.

The beauty of this relationship between the number of photons per square meter per second at each wavelength and the temperature is that it is a single parameter model. That is, we can plug in the temperature of the source and solve for the output at any selected wavelength. This will play an important role in how digital cameras adjust the color response for each photo, which is discussed in Chapter 16.

Fluorescent lights are not so easily characterized since their light is produced in a very different way. For lasers and LEDs, the whole concept does not apply. Color temperature is a shorthand way for stating the relative amount of light being emitted at each of the wavelengths from a source as long as the source behaves reasonably like a black body.

In the light bulb example introduced earlier, the given photon was created by a distinct energy transformation, so the photon has exactly that amount of energy captured in its electromagnetic wave. Photons have no mass, so the energy is not a kinetic energy like that of a bowling ball rolling down an alley. In order for the photon to be absorbed, it must find an atom in which it can effect a transformation of the exact amount of the energy that it carries. If the photon carries X joules, it needs to find a location that will accept X joules—not 0.9 * X joules, and not 1.1 * X joules, but exactly 1.0 * X joules. It is possible to create molecules of materials that have the selected energy bands for electrons and therefore are able to selectively interact with certain wavelengths (colors of light). These molecules are used to make dyes.

In some cases there is a two-stage process. The molecules absorb a certain amount of energy and elevate electrons to an allowable higher energy state, as before. But those electrons drop back to a different lower state. This lower state may not be as stable as the original lower (ground) state; it is called a meta-stable state. The lower state is not the ground state, but is higher than that. This causes the molecule to emit a photon, but since the transition is not as large as the energy absorbed, the photon will have a lower energy and a longer wavelength. In summary, the molecule absorbes a photon at a relatively high energy level—let's say, blue. This imparted energy raises an electron accordingly. Then the electron drops to a new state that has higher energy than the original ground state, but less than the elevated one. The transition of the electron results in the emission of a photon. But since the downward transition involves less energy than the original photon, it has a less energy and a longer wavelength—let's say, orange. The material is said to fluoresce or phosphoresce. In fluorescent materials the electrons fall to a lower state fairly quickly, whereas in phosphorescent materials the time interval is somewhat longer. In both cases, the electrons do not fall all the way back to their ground states, thus the energy released is lower than that absorbed. The electrons in the meta-stable states will ultimately drop back to the ground state, but the energy transitions are small enough that emissions are not visible.

In fluorescent light bulbs, the exciting light is ultraviolet (shorter than blue, which is the shortest wavelength that people can see), and is created by the breakdown of mercury in an electrical arc. The ultraviolet light then strikes the inside of the tube. The light that we see from these tubes comes from the coating of fluorescent dyes on the inside of the glass tube. Several dyes are used to get a color that generally is perceived as "white." Many body fluids also are fluorescent. They tend to absorb ultraviolet (black) light and emit in the range of the green to orange portion of the spectrum.

Photography based upon this effect is referred to as *alternative light source* photography.

When a photon is created, it starts to move along a fixed path at a very high velocity, the speed of light, c. This velocity—300,000 meters per second in a vacuum—is very fast! It is so fast, in fact, that nothing can move faster than this: not matter, not energy, not information, nothing. However, when light moves through a medium other than a vacuum, it moves more slowly. Air is a medium that is very close to a vacuum with respect to the velocity of light, so the reduction in velocity is generally negligible. However if the air is more dense (cold), the velocity will be slightly lower. When light shines through air that has significant local variations in temperature, the objects that are seen through it appear to shimmer. In solid materials, generally the effect is significantly greater. Glass for example might slow the light by as much as 20% or so, depending on specific composition. The velocity of light in a vacuum divided by its velocity within the medium is called the *index of refraction*. The larger the index, the lower the velocity in that material.

When light moves from one medium to another a number of changes to the light beam occur at the boundary. As indicated previously, the light slows down. If the incoming beam is at some angle other than perpendicular to the surface, it will change direction as it enters the new medium. Also, some of the light will be reflected back off the front surface. These three changes will depend on the differences in the refractive indices of the two materials. The light beam will also be chromatically dispersed. Blue light changes more sharply than green, which changes more than red. The angle at which photons' paths are altered depends upon their wavelengths. This is to say that the refractive index of a material is different for the different colors. The degree to which the beam diverges is measured by the Abbe number. The refractive index and the Abbe number are somewhat independent of each other. So a glass can be formulated with a preselected refractive index and a preselected Abbe number. Figure 3.3 shows some values for different glasses taken from specifications in advertising literature.

The bending of a light beam by means of refraction is one of only a few ways to effect change. Reflection, as off a mirror, is another. Diffraction is the bending of a light beam as it passes close to the edge of a mechanical member, such as the metal ring that holds a lens together and provides for mounting. Light beams can also appear to be bent by the distortion of space due to the gravity of massive bodies such as stars, but this is an issue for astronomers and not forensic photographers and examiners.

THE OBJECT

When we take a photograph, we take a picture of something. That something is the object. Usually the object comprises several items that exist in three

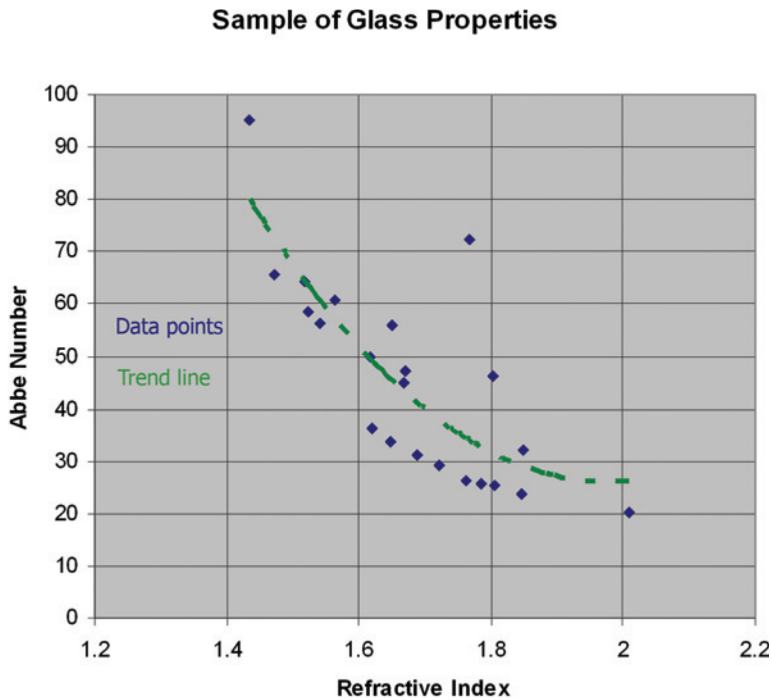


FIGURE 3.3 *Sample of Glass Properties. The data points are for individual formulations of glass, and their positioning shows both the Abbe number and the refractive index. The lens designer has many options in selecting types of glass for a given lens.*

dimensions. When taking a photo of a crack in a lamp on a table at a crime scene, the object will be comprised of the lamp, the table, possibly the lamp shade, other items on the table, and the wall behind. The key subject of the photo, however, is the crack in the lamp base. The light coming onto the subject comes from at least one source, so it arrives from at least one direction. Some of the items in the object will cast shadows on other items. Some of the items will be made of a dark material and some will be light in color. If the light falls directly on a light item, it will be much brighter than another item that is both dark in color and in a shadow. Some of the items are closer to the camera than the key subject and are said to be in a *prefocal area*. Other items are further away and are *postfocal*. The photographer must consider this complex array and make decisions about how to take the photo.

In commercial studio work, the photographer has significant control over the composition and lighting, but in crime scene work there is no option but to leave the items as they are. The lighting and camera angle can be adjusted to some degree, but not nearly as much as in the studio. Only the camera settings can be adjusted fully. The human visual system automatically and subconsciously makes adjustments for all this complexity, so we tend to be unaware of it, or not pay any attention to the issues. But mechanical devices like cameras cannot do this. One additional factor to consider about objects is that the items that are very far from the camera will appear small compared to their true size. If the same object were closer, it would appear larger. It is a linear function. Move the item twice as far away and it seems to be half its size. Move it infinitely far away and it becomes a dot, located at what is known as a *vanishing point*. Visually, people see this effect and expect it. Most lenses that capture images over a wide angle, so-called fish-eye lenses, will appear to be distorting the image. In reality they are making the items that are far away appropriately smaller than those that are close. The photographer must be aware of these issues and take a set of photographs that help the viewer see the scene as it was, see the details that are important to understanding what happened, and not exaggerate or mislead the viewers.

INTRODUCTION TO LENSES

We have all worked with a magnifying glass at one time or another, and know that these are sometimes called "burning glasses." The latter designation gets to the heart of the matter. The burning glass, which of course is a lens, takes the beam of light from the sun and causes it to converge onto a small spot. In this way the energy from all the photons in the beam is concentrated in a single, small location, and is able to heat that spot to the point where it starts to burn. There are two key functions involved: (1) collect the light rays, and (2) redirect them in a predictable way. Camera lenses perform the same two basic functions in cameras. They gather light rays from the scene, or object, and redirect them to create an image of the object (without starting fires).

A lens comprising a single piece of glass is a simple lens. There are two fundamental types of simple lenses commonly used in cameras, shown in Figure 3.4. One type is convex toward the center and is called a convex or positive lens. The other type is concave toward the center and is called a concave or negative lens. Convex lenses cause incoming light rays to bend toward each other, and under the right conditions, converge to a point. At that point the rays will cross each other and then diverge beyond that distance. A concave lens causes the incoming light rays to bend away from each other and diverge forever.

Simple lenses can be used in combination with each other to comprise a compound lens. The lenses can be physically separate from each other, or they can be cemented together. For example, with careful design of the surface curvature and careful selection of glass properties (index of refraction and Abbe number) a convex lens can be mated with a concave lens to minimize chromatic dispersion. Essentially all camera lenses are compound lenses. So are the lenses in telescopes and microscopes. Eyeglasses and magnifying glasses are simple lenses. With compound lenses, it is possible to design for a wide range of special properties and very high performance. Simple lenses are usually only functional over limited ranges.

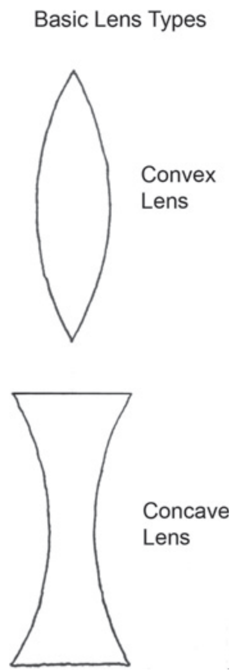


FIGURE 3.4 *Basic Lens Types. The upper drawing shows the profile of a convex lens, the kind that is used in magnifying glasses. The lower drawing depicts a concave lens, which does not cause the creation of a real image the way a convex lens does. Convex and concave lenses often are used in combination to control chromatic dispersion.*

In addition to different types of lenses, it is important to recognize that there are two types of images: real and virtual. If a surface is put in the right place, an image that can be made to appear on that surface is a real image. A common example is the image you see on a movie screen. An image that can be seen in space, but is not projected onto a surface, is a virtual image. Camera lenses and magnifying glasses make real images whereas microscopes and telescopes make virtual images. In order to see the images from microscopes and telescopes, it is necessary to add another lens and make the combination produce a real image. If you look through the telescope, your eye lens is the additional lens, and what is produced is a real image on the retina on the inside of your eye. Another option is to add a camera and use the camera's lens to create a real image on the sensor chip or the film in the camera. Your eye requires that a real image form on the retina in order for you to see the content of the image. Likewise, a camera requires that a real image be formed on the sensor.

Earlier in this chapter it was pointed out that when light passes from one medium to another one with a different refractive index, the light beam is redirected. A light beam comprises a flotilla of photons, and each photon has a path called a *ray*. So, when light passes from air into a lens, striking the surface at an oblique angle, the light rays are bent to an angle that

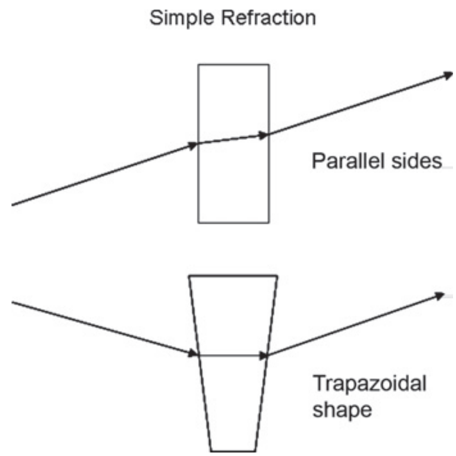


FIGURE 3.5 *Simple Refraction. Two profiles for blocks of glass are shown. One has parallel sides and the other has them at some oblique angle. Both are shown with a ray of light traversing, indicating how the ray will bend when going through.*

is less oblique. When the light comes out the other side, it will bend once again, becoming more oblique. A tracing of the rays shows the phenomenon in Figure 3.5. In the figure, one of the pieces of glass is flat and the other is trapezoidal. The degree to which the rays will be bent depends upon the shift of the refractive index.

Consider the trapezoidal cross-section. Assume that we had a series of concentric glass rings, as shown in Figure 3.6. The rings each have a trapezoidal cross-section, and the angles increase as they progress from the center of the circle out to the edges. If the angles for the series of rings were selected properly, then each ring would take light coming from a source on the axis of the system and redirect it so that they all recombine at another point on the axis on the other side of the system. Taking this a step further, assume that all the trapezoidal rings are very thin and that they are cemented together. The result is a simple lens.

Now take the source of light and move it extremely far away but still on the axis. It approaches becoming a point of light, like a star in the sky. For all intents and purposes the rays from such a source that fall within the area of a camera-sized lens are all parallel to each other. The lens will bend all the rays and cause them to converge at a point. The distance from that point, along the axis, to the center of the lens is the *focal length* of the lens, F. If the source is moved in along the axis, the point of convergence will move further away from the lens. When the source is at 2F, the convergence point will be at 2F on the other side. As the source moves from 2F toward F, the convergence point moves further away, eventually becoming infinitely far away such that the outgoing beam has parallel rays. If the source is moved even closer, the rays will no longer converge.

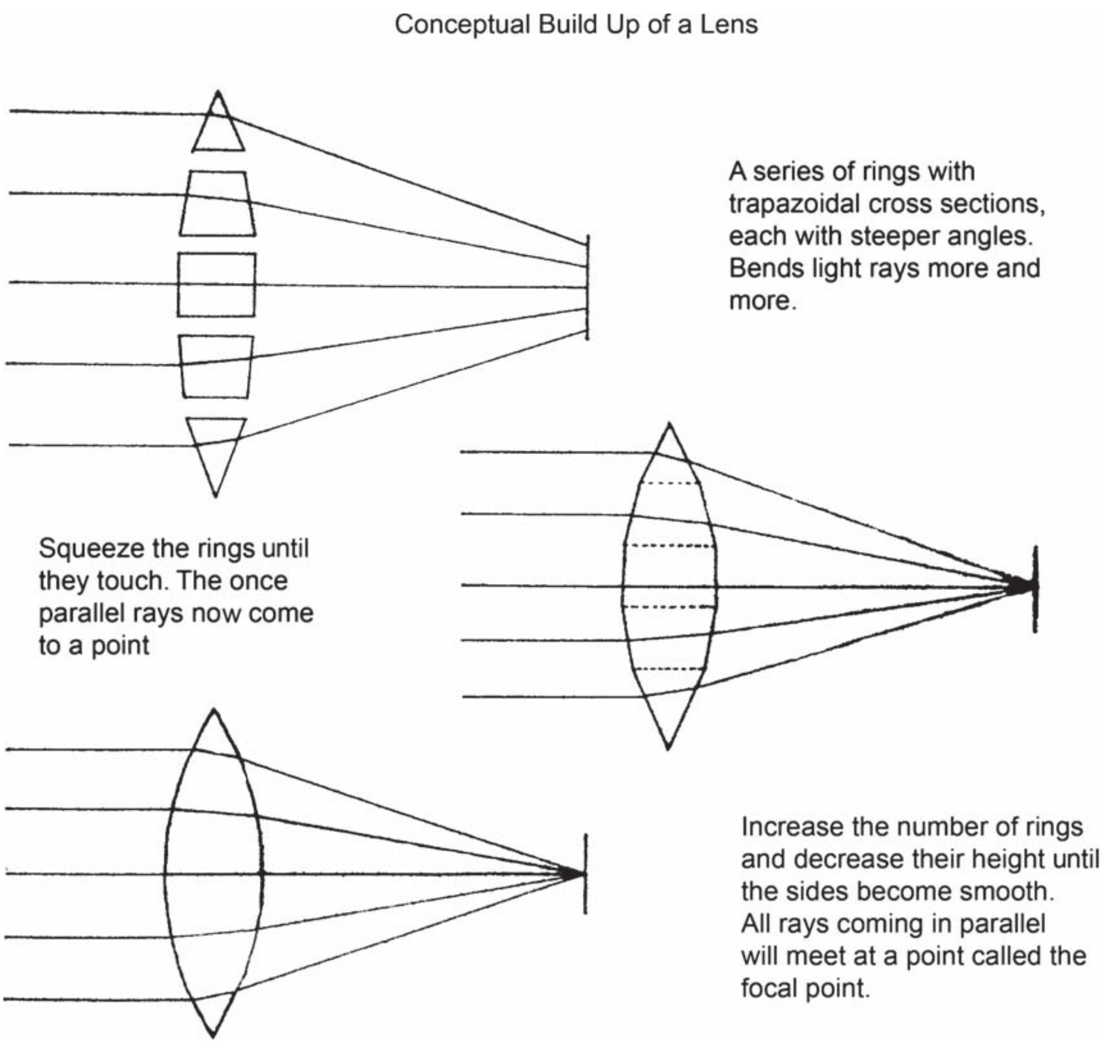


FIGURE 3.6 *Conceptual Build-up of a Lens. Shows how the basic refraction shown in Figure 3.5 is applied to the creation of the curved shape associated with a convex lens.*

Going back to having the source a finite distance from the lens and beyond F, move the source off the axis, say to a location above the axis in the drawing in Figure 3.7. The point of convergence will move to a point below the axis but will still be in the same plane as it was when the source was on the axis. Now consider that an object is, in essence, a series of tiny sources. Each item has a surface, and light is reflecting off those surfaces so as to make them similar to very small light sources. Some of the light from all these points is captured by the lens. If the object were flat, further away than 2F, and perpendicular to the axis, all the object points would converge to points at some plane on the other side of the lens and create a real image of the object. The image will be smaller than the object and it will be upside down. If the object is moved further away, the image will become proportionately

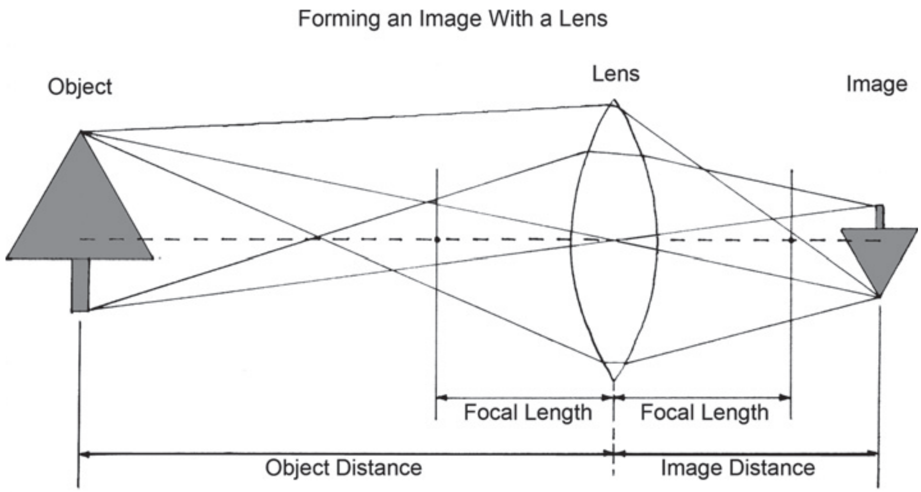


FIGURE 3.7 *Formation of an Image with a Lens. The figure shows a lens in profile and an object on the left. Light rays from the various points on the object (two points are shown) go through the lens and converge on the far side, creating a real (projectable) image. Note that the focal length of the lens is shown. The object is well beyond the focal point on its side and the image is formed between one and two times the focal length on its side.*

smaller. If it is moved closer than 2F, the image will become larger than the object. Positive lenses, be they simple or compound, are used to form real images.

LENS PROPERTIES AND MEASUREMENTS

Modern lens makers have been able to achieve very high quality devices at surprisingly low cost compared to just a few decades ago. And, with all the improvements in both film and digital sensors, all the advances can be appreciated by the photographer. There are four key lens properties that will be discussed here since they are the most applicable to the broad range of photographic applications. Those properties are:

1. Focal length, which determines the camera-to-key subject distances that can be used.
2. Speed, which is controlled by the diameter of the lens and governs the range of lighting conditions over which the camera can be used:
   - f/stops, which are the light control settings used in cameras.
   - t stops, which are the overall indicators of the light passing through the lens.
3. Resolution, which refers to the degree to which a lens can take sharp pictures.

4 Aberrations and distortions:
- Chromatic aberration, which refers to which light of all the wavelengths comes into sharp focus in the same plane.
- Spherical aberration, which refers to the degree to which light from a wide range of input angles comes together in the same plane.
- Pin cushion and barrel distortion, which cause the image to be warped.
- Flare, which is the random scattering of some of the light going through a lens.
- Diffraction, which is the unintended bending of some light rays due to the physical housing of the lens.

Focal Length

In an earlier section, the basic concept of focal length was described as the distance from the lens to the plane where light from parallel light rays will come to a point. It is traditionally measured in millimeters. All simple lenses have a fixed focal length, which is a direct result of the way the lens was ground. It can be used to create images of objects over a wide range of distances, but the distances in conjunction with the focal length will determine the size of the image. The focal length will also determine the angle within which the object must sit in order to be in an image frame of a given size.

In a camera, the frame size and shape of the sensor are fixed by the design of the camera. If it is a 35 mm film camera, the frame is 36 by 24 mm. Digital cameras vary somewhat, but a nominal size might be 24 by 18 mm. These frames have diagonals of 43 mm and 30 mm, respectively. When the focal length of the lens is equal to the diagonal of the image frame, that lens is called a *normal lens*. It turns out that at this length the image will be close to that which a person would see if he or she were standing at the point of the camera. Humans can see across a rather large total angle of view, but only the central 50 degrees yield good detail and rendition. If you extend your arms and hold you hands out on either side of your head, and look straight ahead, you will be aware that something is on either side of you, up to an angle on the order of 170 degrees. But you will not be able to see any detail. For example, you will not be able to distinguish fingers or if anything is being held in the hands. As you swing your arms forward, you will reach a point at which you can see fingers and clutched items clearly. If you were to measure the angle formed by your arms, it would be approximately 50 degrees. Long ago it was noticed that when the focal length of a lens is about equal to the diagonal of the image frame, the field of view is the same as that you would visually see. Hence it is called the normal lens. This is not a highly precise measurement, but it tends to work well over a reasonable range. Some 35 mm film cameras are equipped with a 50 mm lens, which is called normal, and others with a 40 mm lens also called normal. Both work well enough.

If the focal length of a lens is longer than the normal length, then the lens is called a telephoto lens. If it is shorter than normal, it is called a wide-angle lens. Table 3.1 shows these relationships numerically, indicating the key lens relationships of focal length, focus, and magnification. The table shows, for several image frame formats, a number of focal lengths and how those would be classified. It also shows a column for relative magnification. This is not absolute magnification but the degree to which each lens will magnify objects compared to the normal lens. Bigger numbers indicate that an image of the same object at the same distance would be continually larger as the focal length is increased. Absolute magnification is the size of the image divided by the size of the object. Longer (focal length) lenses magnify more and cover a smaller angle of view than shorter ones. Simplistically, it is a linear relationship: if the focal length is doubled, horizontal and vertical distances covered within the frame will each be cut in half, and the area covered will be reduced by a factor of four.

In mechanical cameras, the image frame is fixed and the assembly that holds the lens is fixed relative to the plane of the image frame. If the key subject is to be brought into sharp focus, however, the distance between the lens and the image plane must increase as the key subject is placed closer to the camera. To accommodate a range of key subject distances, the mechanism holding the glass elements must be able to move relative to its mounting on the camera. When we change this location, we are adjusting the focus, or focusing the camera. Focusing the camera does not involve changing the focal length!

Note that some of the items in the object are not in the same plane as the key subject. Some are between the key subject and the camera and others are beyond it. In general, some but not all of these items will be in good focus. The closer they are to the plane of the key subject, the more sharply they will be rendered. In general, there is a range of distances in front of (prefocal) and beyond (postfocal) the key subject at which items will be relatively well rendered. The range of relatively good focus is called the *depth of field*, or *depth of focus* (technically these terms are different, but in common practice they are used interchangeably).

Most camera lenses are not simple lenses, but rather are compound lenses with several elements. It is important to note that the focal length of a compound lens is related to both the surface grind of the glass elements and the spacing between the elements. Figure 3.8 shows the basic relationships associated with a compound lens comprising only two elements. Note that the focal length is no longer controlled only by the way the glass elements are ground, but is now also dependent on the distance between those two elements. If the device that holds the glass elements has the ability to move the glasses relative to each other, that mechanism can adjust the focal length of a mounted lens. This cannot be done with a simple lens. When a

Table 3.1 *Camera Lens Categories*

Lens Descriptor	Angle of View Degrees	Relative Magnification	Common Camera Formats 4 × 5 in 102 × 127 162.6	6 × 7 cm 60 × 70 92.2	35 mm 24 × 36 43.3	digital* 18 × 24 30.0	Format name Dimensions (mm) Diagonal (mm)
Extreme	120	0.25	49	28	13	9	focal length (mm)
Wide	100	0.40	64	36	17	12	focal length (mm)
Angle	90	0.50	79	44	21	15	focal length (mm)
Moderate	88	0.51	83	47	22	15	focal length (mm)
Wide	70	0.67	110	62	29	20	focal length (mm)
Angle	60	0.89	140	79	37	26	focal length (mm)
	59	0.90	144	81	38	27	focal length (mm)
Normal	50	1.00	163	92	43	30	focal length (mm)
	45	1.20	190	107	50	35	focal length (mm)
Moderate	42	1.70	209	118	55	39	focal length (mm)
Telephoto	27	2.30	380	214	100	70	focal length (mm)
	12	4.50	758	428	200	139	focal length (mm)
Extreme	12	4.50	758	428	200	139	focal length (mm)
Telephoto	8	6.80	1137	642	300	209	focal length (mm)
	6	9.00	1516	856	400	279	focal length (mm)

**There are several digital still camera formats and this one is merely an indicator*

compound lens has the ability to have its focal length changed, it is called a *zoom lens*. Usually, zoom lenses are designed such that changing the focal length—that is, the level of magnification—will not change the focus setting. Thus the photographer can focus the lens first and then adjust the level of zoom and not have to keep cycling back and forth to get the proper photo.

Lens Speed

In the earlier discussions of sensitometry and dynamic range, the major input variable was exposure, or Log exposure. Photographic sensors respond to the number of photons per unit area that impinge upon them. There are two factors involved in determining this: (1) the spatial concentration of photons and (2) the length of time during which they impinge. The first is the number of photons landing on a unit of area in a unit of time (lux), and

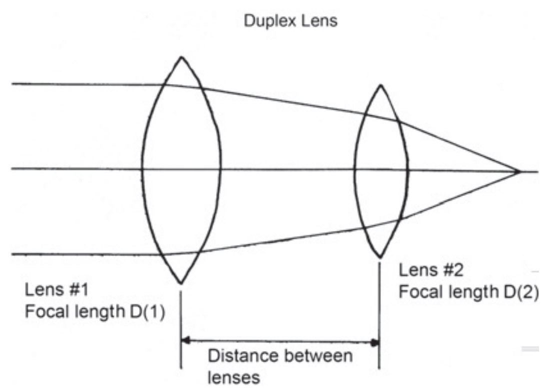


FIGURE 3.8 *Duplex Lens. Shown is a pair of convex lenses in a single mount. Although the focal lengths of both lenses are important to knowing the equivalent focal length of the combination, the distance between them is also important. The individual focal lengths are ground into the glass and cannot be changed. The distance between them can, thus allowing for creation of a zoom lens.*

the second is the duration of the exposure, or exposure time (seconds). In specific terms, this could be the number of photons per square meter per second and the number of seconds of exposure. The result of multiplying these two terms is photons per square meter; that is, lux × seconds = lux-seconds, which is the most common (currently) unit of measure for exposure. The lens does not usually do anything to control the time of exposure, but it has a lot to do with the lux reaching the sensor.

Since the basic measure of lux is photons per unit area per second, then the number of units of area of the lens opening will directly control this factor. If the area of the opening of the lens is doubled, then the photons passing though it (all other things being equal), will be doubled as well. Most camera lenses contain a device that can be adjusted to change the effective open area of the lens. It is called an iris diaphragm, or simply a diaphragm. In the human eye, there is an iris as well, and it serves the same purpose. When it is dark, the diaphragm opens, the pupils are said to be dilated, and more of the available light is allowed to enter the eye. When it is bright, the diaphragm closes down somewhat and less of the light enters the eye. The diaphragm in a camera is usually round, just as in the eye, and so a way to measure the opening is to measure the diameter. But there is a quadratic relationship between the diameter of the lens and its area—the area is equal to pi ($\pi = 3.14159$) times the radius (half the diameter) squared, or $A = \pi * r^2$.

The glass elements in a compound lens have fixed diameters, and so they have fixed open areas. Any diaphragm that will be used to change the size of the opening can only decrease it. Basically, what is used is a series of overlapping leaves mounted with pivots that are on a circular ring. As the leaves are rotated, they either decrease the size of the opening they establish in the center of the ring, or open it back up. So in a practical lens, we start with a maximum opening and then there are means to systematically decrease it. Because of the logarithmic behavior of human vision and photographic systems designed to serve as its surrogate, the settings on the control iris should follow a multiplicative series. The one used is in sequential factors of two. So when fully open, the area might be called one unit. The next setting would be one half that unit. The third would be one-fourth, then one-eighth, one-sixteenth, and so on. Considering the formula for the area of a circle, the series of circles with areas sequentially decreasing systematically by factors of two will result in a series of diameters that change by sequential multiples of the reciprocal of the square root of 2, or 1.414.

The preferred way to adjust a lens is to reset the area by successive factors of two, or go to openings that have sequential diameters decreasing by $1\sqrt{2}$. So, either the diameter of the area can be used to designate the various setting or some other similar metric must be developed. As it turns out there is a better metric to use: the *f/stop* (pronounced, "f stop"). The defining equation for the f/stop is

$$\text{f/stop} = \text{focal length} \div \text{diameter of the opening}$$

Table 3.2 shows the area, diameter, and f/stop for openings. These are in order of decreasing areas, where the rate of decrease is successive factors of two. Note the listing of f/stop values. As the f/stop number gets larger, the area gets smaller and the series of numbers is a bit peculiar. The sequence of numbers is well known to many photographers, but the source of them is often a mystery. If you read the mystery novel *The Da Vinci Code*, by Dan Brown, that particular series of numbers played a key role in unraveling one aspect of the mystery. Getting back to photography the question remains, why use f/stop values instead of diameters or areas? The answer is because f/stops are consistent across lenses.

Consider using two different lenses in the same photographic setup: a 100 mm lens and a 200 mm lens. From the material discussed earlier, the 200 mm lens will cover one quarter of the object compared to the 100 mm lens. The object is the same and its lighting is the same, so the number of photons per square meter coming from the object will not change. But the 200 mm lens, since it sees only one quarter of what the other lens sees, will capture only those photons from that one quarter. Thus the number of photons going to the 200 mm lens is only one-fourth compared to the 100 mm lens. But both lenses will each take the photons captured and spread them across the same image frame area. So the 200 mm lens will impinge fewer photons per square meter than the 100 mm lens—one fourth as many, to be exact.

Table 3.2 *f/stop Parameters*

Focal Length, L		100 mm lens		Focal Length, L		200 mm lens	
Area ratio	**Area of opening**	**Diameter**	**f/stop**	**Area ratio**	**Area of opening**	**Diameter**	**f/stop**
1.0000	7854.0 sq mm	100.00	1.00	1.0000	31,415.9 sq mm	200.00	1.00
0.5000	3927.0 sq mm	70.71	1.41	0.5000	15,708.0 sq mm	141.42	1.41
0.2500	1963.5 sq mm	50.00	2.00	0.2500	7854.0 sq mm	100.00	2.00
0.1250	981.7 sq mm	35.36	2.83	0.1250	3927.0 sq mm	70.71	2.83
0.0625	490.9 sq mm	25.00	4.00	0.0625	1963.5 sq mm	50.00	4.00
0.0313	245.4 sq mm	17.68	5.66	0.0313	981.7 sq mm	35.36	5.66
0.0156	122.7 sq mm	12.50	8.00	0.0156	490.9 sq mm	25.00	8.00
0.0078	61.4 sq mm	8.84	11.31	0.0078	245.4 sq mm	17.68	11.31
0.0039	30.7 sq mm	6.25	16.00	0.0039	122.7 sq mm	12.50	16.00
0.0020	15.3 sq mm	4.42	22.63	0.0020	61.4 sq mm	8.84	22.63
0.0010	7.7 sq mm	3.13	32.00	0.0010	30.7 sq mm	6.25	32.00
0.0005	3.8 sq mm	2.21	45.25	0.0005	15.3 sq mm	4.42	45.25
0.0002	1.9 sq mm	1.56	64.00	0.0002	7.7 sq mm	3.13	64.00
$A(j) = A(i) \div 2$	$A = \pi * r^2$	$D = 2 * \sqrt{(A/\pi)}$	$f/stop = L/D$	$A(j) = A(i) \div 2$	$A = \pi * r^2$	$D = 2 * \sqrt{(A/\pi)}$	$f/stop = L/D$

It would be nice to have a control on the lens that allows the photographer to compensate for this change. That is what the f/stop does. The f/stop is the focal length divided by the diameter. If the f/stop is kept constant from one lens to the other, the diameter must increase in proportion to the focal length. And since the open area of the lens varies with the square of the diameter, it will result in a lens opening (in square meters) that is exactly what is needed to compensate for the change in focal length. For this example, the opening of the 200 mm lens set at the same f/stop as that of the 100 mm lens will be the square of 200/100, or a factor of 4—just what is needed. Using the shower analogy (which will be developed in chapter 4), we can say: The area of the opening on the can representing the 200 mm lens is four times as large as the can representing the 100 mm lens when both are at the same f/stop. Nice! It is important to remember that as the f/stop number goes up, the amount of light going through the lens goes down!

When light travels from one medium to another, some of the light reflects off the surface of the new medium. This means that some of the light is not going through the medium. In addition, all media except vacuums absorb some of the light that goes through them. Again, some of the light coming in does not go through. Both of these contribute to a loss that is generally a small fraction of the incoming light and the effect is normally ignored. But if there are several elements, and there is a loss associated with each one, the multiplicative effect (as described by Beer's law) can start to add up and become an issue. For example compare two lenses—one, a basic 50 mm lens, and the other, an extended range zoom lens set to 50 mm. The zoom lens will probably have many more elements than the more basic lens, and even if the two lenses are set to the same f/stop, the exposure with the zoom lens will be lower due to more light being absorbed and reflected back. The discrepancy can be rated as if it were an f/stop, and often it is called a T-stop. More often it is given as a unit of density; typically the density associated with a clear material such as glass or plastic is close to 0.04. If there are three such elements, the density for the combination will be 0.12, which is the sum of the values for the individual elements.

As mentioned earlier, most modern camera lenses have several elements. To reduce the effect of front surface reflection for each element, an antireflection coating is added to the surfaces. Typically this is a coating of a material with an index of refraction that is the geometric mean of the two materials, and the coating is made to a thickness of one-fourth of a wavelength of green light, which is very thin. Green light is in the middle of the visible spectrum and has a nominal wavelength of 550 nanometers. So the coating is about 138 nanometers thick. If you hold a lens in your hand and move it around relative to the ambient lighting, you will find an angle at which you can see color off the surface. This is due to the coating. To preserve this very thin coating, lens surfaces should be treated with extreme care. Keep them clean, and free of dust and fingerprints. If some dust does get on the lens, use a lens brush/blower to clean it. If a fingerprint does get on the lens, use only a lens cleaner to clean it and carefully follow the instructions given with the cleaner.

Resolution

Take a magnifying glass and look closely at a mid-range portion of any of the pictorial images in this or any other book. All of a sudden the area that appeared to consist of continuous coloration now is composed of a series of very fine, separate, and distinct dots. Your eye alone was not able to distinguish separate and distinct items in the object as separate and distinct. When the magnifying glass was added, the ability to resolve the dots was realized. Resolution is the ability to see as separate and distinct items in the image that are in fact separate and distinct in the object. Resolving power is a measure of the level of fineness that can be resolved. It can be applied to a lens, a film, a digital camera, or a complete television system. In most cases, resolution is not a yes/no determination. As the input image becomes finer and finer, the ability of the system to recognize separate items as separate entities decreases. First the space between them becomes blurry, then parts of the items bridge the space, then the items become indistinguishable. The general rule is to find the spacing of items at which there is noticeable bridging and designate it as the resolution limit. It is generally stated as line pairs per millimeter. Figure 3.9 shows the United States Air Force test target that frequently is used to make resolution measurements (this can be downloaded from the Internet; search "USAF test target") as well as a test target that might be used to measure modulation transfer.

Another way to determine the resolution capability of a system is to determine its modulation transfer function, or MTF. To make this measurement, a test target is created that contains a series of patterns that vary in

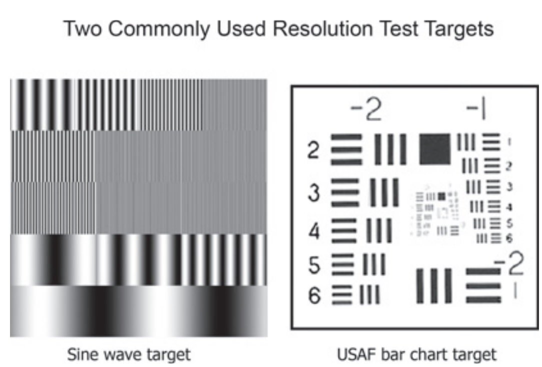


FIGURE 3.9 *Common Resolution Test Targets. Depicted are two common test targets: the three-bar type and the sine-wave type. By knowing the specifics of each target and the photographic setup, one can measure the resolution of a camera or system.*

reflection or transmittance according to a series of sine waves. Different parts of the target have sine waves with different spatial frequencies, measured in (sine wave) cycles per millimeter. In each case the change from the high reflectance (or high transmittance) peaks to the low reflectance (or transmittance) valleys is the same for all frequencies. The system or camera to be tested is used to make an image of the test patterns and then the peak and valley values are measured for each frequency. At the higher frequencies (more cycles per millimeter) the modulation (peak-to-valley ratio) will decrease at an increasing rate. Typically the frequency at which the modulation drops to 50% of the low frequency level is stated as the frequency response of the system. In some cases the system will respond to edges differently than it will to simple sine waves. This is the result of nonlinear edge effects. In such cases the MTF will go above 100% at some points. This will make simple interpretations difficult.

Both Resolving Power and Modulation Transfer Function methods have proved useful for determining performance. MTF has the advantage that the MTF of a system can be calculated from the MTFs for each of the elements, and so it is quite useful in designing systems. But resolving power is easier to measure and the result is more intuitive. In both cases, higher resolving power numbers or higher frequency response cut-off points will be reasonably good predictors of how sharp an image will appear.

Aberrations and Distortions

As with anything else, perfection is impossible. The closest thing to a perfect lens is a diffraction limited lens. When light rays pass close to an opaque member they tend to bend as if they were trying to wrap around that member. The phenomenon is known as *diffraction*. All lenses have a mounting ring and most have a diaphragm. Both of these will cause light rays to experience a bend that is in addition to the bending caused by the glass in the lenses. It is inescapable and so it is impossible to build a lens that performs better than these limitations imparted by the physical structure.

But this is not the only limitation. It is impossible to make glass that is perfectly homogeneous. So there are minute variations in the refractive index in certain spots in the glass. In addition, it is impossible to grind a perfect lens surface, so there will be minute errors in the surface of the glass. These will cause some light rays to stray from where they were designed to go and degrade the image.

The common perception is that lenses are ground with surfaces that are part of a sphere. This is not the case. A purely spherical surface will cause the rays that are further from the center to bend too much compared to those nearer the center. If all the rays do not come together in the same plane, the image will be blurred. This is called *spherical aberration*. It is important that a proper grind be accomplished in order to keep the image sharp. Spherical aberration will lead to low resolution.

The bending of light rays by a lens is the result of refraction. But the refractive index is slightly different for photons of different wavelengths. The result is that a simple lens will cause different colors to focus in different planes. Images of sharp-edged items would appear to have color fringes surrounding them. This is called chromatic aberration. This problem is usually solved by making doublets. These are two lens elements cemented together. One is a convex lens and the other is a concave lens. The refractive indices and the curvatures of the two are selected so that they do not fully cancel each other out and on balance remain effectively convex. However the Abbe numbers for the two are selected so that the refractive index shifts with wavelength are largely canceled out. The process is well enough understood and the range of materials available today is sufficient to make chromatic aberration a very minimal issue with most camera lenses.

If you take a photo and the key subject is a flat, square item there are three common shapes that the image may take. The image might show a square, which is, of course what is sought. Or, the image might be almost a square, but the sides bulge outward, resembling a barrel-like figure. This is called *barrel distortion*. The third option is where the sides are pulled inward. This is called *pincushion distortion*. Both distortions are due to design flaws.

Some lenses have a very short focal length relative to the normal lens. These are sometimes called *fish-eye* lenses because the image looks like what we imagine a fish would see. This is not barrel distortion. With wide-angle lenses the object items well off to the sides are considerably further away from the lens than the ones near the center of the frame. Since they are further away, the rules of perspective dictate that they will appear smaller in comparison to those that are closer. The result is the apparent deformation of the image. Some wide-angle lenses are designed to correct for this effect. The opposite occurs with long telephoto lenses. Since things are magnified more than with the normal lens, items that are at different distances from the camera appear to be too close to each other compared to what a person would see if looking at the object. This is called depth compression and it is not really a distortion.

It has been pointed out that not all of the rays that go through a lens go exactly where they are supposed to go. There are reflections off of front surfaces, there are reflections off of support and diaphragm members, and there are errors due to diffraction. All of these might be considered stray light. This stray light is called flare, or flare light. The result is that if the object has an item with a density of 4.0 and another item with a density of 0.04, the image will not show this same range of light levels. The item with a density of 4.0 will have a density of less than that. If the flare is 0.1%, then the maximum density any item can have will approach 3.0 (that is 10^{-3}).

It should be clear from this discussion of distortions, aberrations, and imperfections that all these contribute to loss of resolution. And once these deficiencies have been designed into the lens or manufactured into the lens, there is not much that can be done to circumvent the limitations. If you are seeking to get high-quality images, you will have to invest in high-quality lenses.

Optimal f/stop

The discussion of f/stops so far has dealt with the amount of light passing through the lens. This is not the only factor affected by adjusting the diaphragm, however. As the diameter of the lens opening is increased, the number of rays increases and the angle at which the peripheral rays intersect increases, as shown in Figure 3.10. This means that the angle at which the rays converge on the focal plane is increased as well. In the figure, the disk that results from forming an image of a point is shown as the point is moved from out of focus, into focus, and back out the other side. At first the disk becomes smaller, then the disk stays essentially the same, and finally it increases again. Within the range where the disk does not significantly change in size, the sharpness of the image does not change either—all the associated object distance points appear to be in focus. The range is referred to as the depth of focus, or depth of field. To have a larger opening, which means lower f/stop numbers, rays converging at sharper angles have a smaller depth of field. So, the f/stop setting controls both the amount of light passing through the lens *and* the depth of field.

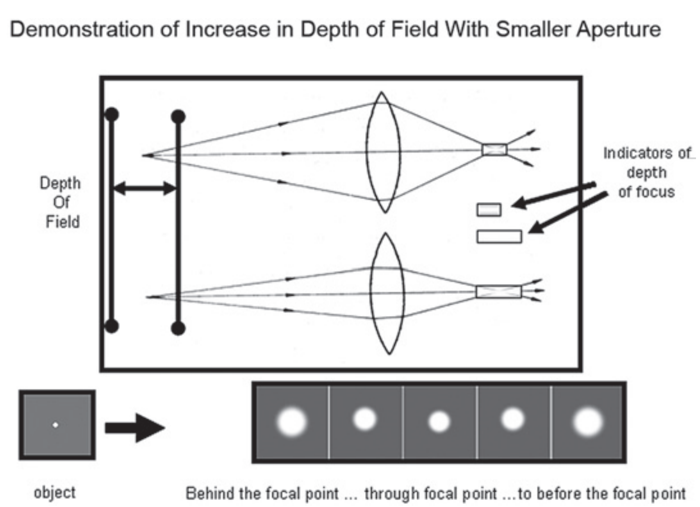


FIGURE 3.10 *Demonstration of Depth of Field Effect of f/stop. The lens in the upper diagram has a wide opening (low f/#) and the one below has a narrow opening (high f/#). At the bottom left is the object, a small spot of light. To the right are the images that would be attained at different distances. In the center is perfect focus, which still has some spreading of the spot. Adjacent to either side are spots at small distance from perfect focus. Beyond this are images from greater distances from perfect focus. The rectangles associated with each lens indicate that the wide open lens rays will diverge faster than those of the stopped down lens, causing the image to degrade at shorter distances from perfect focus.*

Another factor is resolution. According to statistics, if the image is comprised of a larger number of rays, the error associated with creating the sample is reduced (the standard deviation of a sample decreases with the square root of the sample size). This means that one would expect larger openings to be capable of higher resolution. The formulae describing diffraction limitations also show that the resolution capability of a lens goes up with its diameter. These, in fact, are realized, but only up to a point. As the opening approaches the lower end of the f/stop number adjustment range there are very large openings. The rate at which distortions, aberrations, and defects appear increases with the area. These combine to counteract the benefits of more rays converging. As a rule of thumb, the optimal setting for the f/stop adjustment is between one-half to two-thirds the way from the largest f/stop number to the smallest.

The result is that setting the f/stop is a complex issue. First of all it must satisfy the need for proper exposure level. Second, it must satisfy the depth of field requirement, and finally, the resolution requirement should be considered.

LENS FORMULAE

Figure 3.7 shows a simple lens creating an image of an object. The focal length, F, is shown along with the object and image distances. For the simple lens, the basic formula describing performance is:

$$1/F = 1/d_o + 1/d_i$$

where F is the focal length and d_o and d_i are the distances from the lens to the object and to the image, respectively. Note that this supports the assertions made earlier. For example this indicates that when d_o and d_i are both equal to each other, they will be equal to 2F. As the object is moved further away—that is, d_o is made larger than 2F—d_i becomes smaller. When d_o is moved to infinity, the first term becomes zero and d_i becomes equal to F. In the case of the simple lens, the system is symmetrical; if the lens were turned around, the behavior would still be the same.

In the case of a two-element compound lens the formula changes and becomes:

$$F = \{F(1) * F(2)\}/\{F(1) + F(2) - D\},$$

where

$$D < F(1), F(2)$$

The important differences are that the focal length of the combination is dependent upon the distance between the two glass elements, and if the two lens elements have different native focal lengths (F(1) does not equal F(2)) then the system will not be symmetrical. The ability to change the focal length of the compound lens by changing the distance is what gives rise to the ability to make zoom lenses. The fact that the system is asymmetrical gives rise to the fact that the view through camera lenses is different depending upon which end you look through.

Referring back to Figure 3.13, note that the ray going though the center of the lens essentially does not deviate from a straight line. The result is that the triangle created on the object side is similar (the angles are the same and the sides are in equal proportion to their counterparts) to that on the image side. From the geometry of the system, if we know the focal length of the lens, the distance from the key subject to the lens, and the size of the key subject, we can determine the distance to the image, the size of the image, and the ratio of the size of the image to the object, or the magnification. The equation for the magnification is simply:

$$M = d_i/d_o, \text{ or } M = s_i/s_o$$

where s_i and s_o are the sizes of the image and the corresponding key subject. Note that if the key subject fills the entire frame width, then these measures are the image and object frame widths at the points of focus. Combining this equation with the basic lens equation, we find that:

$$s_o = s_i * (d_o - F)/F, \text{ or } w_o = w_i * (d_o - F)/F$$

where the w values indicate object width and frame width, respectively. Restating these relationships it can be shown that the magnification can be determined by knowing the focal length and the distance to the key subject. The equation is:

$$M = s_i/s_o = F/(d_o - F)$$

These equations and various combinations and restatements can be used to determine the photographic setups needed for a number of assignments. For example, if photographing a fingerprint where it is important to know the size that the image represents, we can use the magnification as a function of focal length and distance. With information on the object width and the object distance and the focal length of the lens, we can calculate the angle that the lens will capture. Or if a certain distance and angle are known, we can solve for the focal length of the required lens. With a fish-eye lens, where there are object items that are way off to the side of the object frame, the distance of the side items will be significantly different from that of items directly in front of the lens. We can use the preceding equations to determine the magnification at these different points.

Photometry

In earlier chapters it was established that light is composed of particles of light called photons. Each photon comprises a wave packet, which is a finite length of electromagnetic wave with a dominant wavelength. The wavelength and frequency are inextricably linked according to the simple relationship:

$$\lambda = c/\eta$$

where λ is the wavelength, η is the frequency, and c is the velocity of light. Another important relationship establishes the energy associated with each photon. This is given by:

$$e = h \star c/\lambda$$

or from above,

$$e = h \star \eta$$

where h is Plank's constant, which is equal to 6.626 exp -34 joule seconds. On the right side of the equation we have joule seconds multiplied by the frequency, which has the dimensions of per-second. The product expresses the energy of a photon in joules.

In many engineering and scientific studies, the energy conveyed by a beam of light is often of great importance. The energy of the beam is equal to the sum of the energies of the photons in that beam. The amount of energy imparted to any absorber will show up as a change in temperature, or, in microscopic studies, as transition inside atoms. But, in most visual and photographic applications, where the effect of interest is the liberation of electrons, and the wavelength of the receptors is relatively narrow, the number of photons arriving per second per unit of area is really all that matters. The range of wavelengths in the beam is often confined to groups consistent with the sensitivities of the eye or photographic sensors.

Because of the dichotomy between the needs of the scientific studies and the photographic and visual studies, there are two parallel sets of measurements.

One is used in scientific applications, and that set of measurements carries the additional label *radiant*. The study of these relationships is known as *radiometry*. In photographic and visual applications, the measurements are given the label of *luminant* or *luminous*, and the field is known as *photometry*. Since this book deals with photography, all light measurements will be from the photometric set unless otherwise indicated. This means that the spectral energy of the source is known and well defined, and that the sensor has a response that is tailored to that of the human eye and its constituents. So a light meter designed for photography will read in lux, which is a photometric unit of measure, and it must have a spectral sensitivity comparable to that of the human eye. The same will hold for some lux-based units such as exposure values.

The full set of measurements is large and the relationships are complex. It helps to know that some of them deal with the source of the light (candela, or cd); some with the flow of the light (lumens); and others with the arrival of the light at some surface (lux). And finally there are measures that deal with light reflecting off a surface (candela per square meter). The result is that it is hard to grasp the concepts associated with each of the measures. To help understand the concepts, we will develop the "shower analogy."

THE SHOWER ANALOGY

Image that you have a shower head that sticks out from the ceiling and spurts out water in a wide pattern, as shown in the diagram in Figure 4.1. We will assume that there are a very large number of holes in the shower head to allow

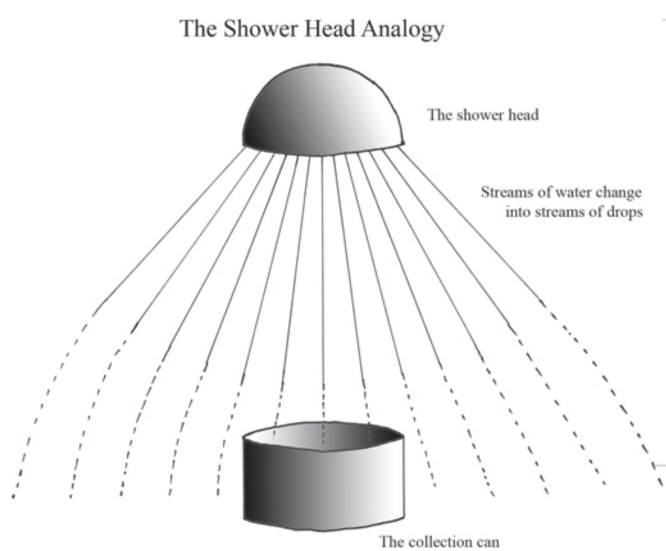


FIGURE 4.1 *The Shower Head Analogy. The shower head spews out diverging water streams that change into drops. The drops are meant to be analogous to photons. The can at the bottom is meant to be analogous to the camera or its sensor.*

for very many streams of water to emerge. We will also assume that almost immediately after emerging, the water in the streams breaks up into droplets. The shower head stands for the light source in our analogy. The droplets of water represent photons. The rate at which droplets are emitted is analogous to candela in a light source.

The streams of droplets diverge from one another as they extend away from the shower head. Once they emerge, they continue on their paths. In the case of photons, they continue to travel in straight lines until something like a lens or a mirror causes them to bend. Unfortunately the droplets from the shower are under the influence of gravity and their paths do not stay in straight lines. However they do continue to diverge. The envelope of streams of droplets is analogous to the conveyance of light from the source to the sensor. The number of droplets passing through a unit of area on the way down from the source is analogous to lumens.

On the floor of the shower there are a number of cans with different diameters placed near each other in the center of the water pattern. Each can has a lid that we can place over its open top to block droplets from falling into the can. The can is analogous to our sensor.

Considering the shower head, it can be characterized by the number of drops coming out per second and by the number of droplets coming out per second per unit area on its surface. The same concepts apply to a light source. It can be characterized by the total number of photons emerging per second, and the number emerging per second per unit area. In the following tabulation are photometric units of measure that that describe a light source. The first, the *candela* (abbreviated *cd*) measures *luminous intensity* and indicates the total number of photons per second emerging from the source. The second measures *luminance* and is given in *cd per square meter* on the surface of the source.

In the shower head analogy, we can increase the intensity by turning up the faucet. This will produce more droplets per second and more per unit area. If some of the holes in the shower head happen to be partially blocked, those areas will have fewer droplets per second per unit area, analogous to reduced luminous intensity.

Term	Description	Measurement	Photon Basis
Luminous Intensity	An indication of the brightness of a light source	candela (cd)	Total number of photons per unit time from a light source
Luminance	The brightness of a surface at a point	candela meter^{-2} (cd m^{-2})	Photons per unit time per unit of area on the surface of a source

Moving on from the source there is a basic measure of the light moving away from the source through space. This unit of measure is the *lumen*, which is a measure of *luminous flux*. There are two basic elements to this measure. The first is the amount of light and the second is the measure of space through which the light is flowing. The unit of light is the candela, defined earlier as the total number of photons per second emerging from a source. The unit of space is the sterradian. It is easier to start in two dimensions. Consider a circle with a point in the center. Draw a line segment from the center to the circle's circumference. That line is, by definition, r units long, where r is the radius of the circle. Now draw another line segment such that it touches the circumference at a distance along the circumference that is r units from the first line. The angle formed by the two line segments is a radian. Radians can be converted to degrees by noting that the circumference is equal to $2 * \pi * r$, and since the amount of circumference covered by a single radian is r, there are $2 * \pi$ radians in the total 360 degrees of a circle.

A radian is a plane angle. Its three-dimensional counterpart is a *sterradian*. To get this, take a sphere with a point at the center and draw a line all the way around—like the equator goes around the earth. Now pick a point on this line and draw another circle that crosses the first at a right angle—just as the Greenwich Mean Time meridian crosses the equator. There is now a sphere with two great circles drawn on its surface. For each of the circles, draw another radius line one radian away from one of the points where the two circles cross. Repeat the process again and a fourth point will be located on the surface of the sphere. Connect all the lines on the surface and a closed figure will appear. It is something like a square in that all its sides are the same length and it has four corners that are all right angles. The difference is that the sides are not straight lines, but are arcs on circles. The four radial lines that go from the center to the four corners comprise a solid angle, and this particular solid angle is one sterradian.

If we consider our light source to be a point source (very, very small compared to the distance from it), and if we consider the source to radiate uniformly in all directions (that is, the luminance is constant in all locations), then the light will uniformly fill each sterradian of light with a certain number of photons per second. The measure that results is called the *lumen* and it is a measure of *luminous flux*. If the source does not radiate equally in all directions, the luminous flux will be different in different directions. In the shower analogy, it describes the number of drops per second per unit area where the unit of area increases as the distance below the shower head increases. Projectors are often rated in lumens, the number of photons per second per screen size, where the screen size increases as the distance from the projector is increased. Note that the number of photons per unit area on the screen will decrease, since the same number of photons must cover a larger area.

Term	Description	Measurement	Photon Basis
Luminous flux	The amount of light emitted into space	Lumen (lm) (cd sterradian^{-1})	Photons per unit time emitted into a solid angle of space from a source

So far the measures of the source as well as the space between the source and the surface have been described. What remains is to measure the light falling on a surface. To do this we will refer to the *illuminance*, which is measured in *lux*. In this case what is being measured is essentially the number of photons per second per unit of area on the indicated surface, or in the analogy, the number of droplets falling on the tops of the cans per unit area per second. It indicates how bright the surface will appear.

Term	Description	Measurement	Photon Basis
Illuminance	Amount of light falling Per unit area	Lux (lx)	Photons per unit time per square meter on a receiving surface

One key topic remains: *exposure*. In earlier chapters, sensitometric curves were constructed where the input was log exposure. Exposure is lux times seconds, or lux-seconds.

Term	Description	Measurement	Photon Basis
Exposure	The amount of light reaching a sensor	Lux seconds (lx seconds)	Photons per unit area on a receiving surface

Getting back to the shower, the exposure is the total amount of water that accumulates in the can. If the can is placed on the floor of the shower, then the droplets will have separated from each other quite a bit. Thus the number falling into our can (which has a fixed opening and therefore a fixed open area) per unit of time will be considerably lower than if we held up the same can waist high, or close to the shower head. If we covered the can and turned on the shower no water would collect. Then if we remove the cover for a certain number of seconds, we will collect a certain number of liters of water. If the lid were kept off longer, more liters would be collected. If the can is moved closer to the shower head, and the short time of lid removal was repeated, we would get more liters of water than if the can were on the floor.

The point is that the exposure is analogous to the total amount of liters collected, which is the sum of the droplets captured. If a can with a bigger diameter is used, it will collect more liters per second and thus is more sensitive to droplets per second per unit area since this value is actual area of the can multiplied by the time the can is open. In silver halide photography the way to make a more sensitive film is to make the silver halide crystals larger. In digital cameras the way to increase the sensitivity is to make the individual pixels larger. The increase in sensitivity is directly proportional to the increase in area, which is the square of the linear dimension.

The system of units is complex and generally only people who work with them regularly can keep it all straight. So there are a few simple concepts to keep in mind. First of all, if the source gets brighter, everything else gets brighter as well, and does so in direct proportion. Second, all the units are linear with respect to this basic brightness, but vision and cameras are responsive to equal proportions of increase, which is a logarithmic scale, and not a linear one. Third, you might want to remember where to look up the definitions of the units since various meters and product descriptions frequently refer to them.

Incident Light and Reflected Light

In the discussion of photometry, there were references to the source, the transmission of the light to a surface, and the light impinging on that surface. In photography there may be other intermediate steps to consider as well. In the simple case, for example, of taking a photo outdoors, there is light that comes from the sun and falls onto the object. Then there is reflection from the surfaces of the object into the camera. Finally there is the light that impinges on the sensor in the camera. In normal parlance, the light falling on the object is called *incident* light. The light coming off the object is called reflected light.

If we were to measure the light arriving at the camera, that might be called incident light as well, but it had better be the light arriving at the camera that is coming from the object and not just any light. If it is not from the object, it will not go into the camera and activate the sensor. The incident light is frequently measured in lux and the reflected light is measured in candela per square meter (cd/m^2). Modern cameras have internal light meters that measure the light that actually goes through the camera's lens. This is particularly effective metering. The new cameras can also use this light to set all the camera adjustments to give the best exposure under the conditions. This will be covered later. For now it is important to be careful which light is being measured when setting a camera.

In careful situations, the photographer will measure key portions of the light incident on different parts of the key subject in order to ensure that those areas are all correctly exposed in the camera. Examples of light levels were shown in Figure 2.3.

There are a variety of functions that light meters might be designed to perform. The issues are incident versus reflected light reading, lux versus exposure value readings, and steady state versus impulse reading. Some light meters are designed to be held in front of the object to measure the light coming onto the subject of the photo. Others are meant to be held in front of the camera and measure the light coming to it. As indicated earlier a variant of this is the through-the-lens (TTL) light meter.

When looking at reflected light one needs to be careful with respect to what the light is reflecting from. This is true of TTL meters as well. There are three common settings. One is set to integrate all the light coming from subjects within the frame. This has limited value since the object is usually composed of several subjects, and not all of them are important. To deal with this, there is a center-weighted setting. In this mode, the light coming from the center of the frame is measured and the periphery of the frame is either ignored or given less weight in determining the exposure. Finally there is the spot meter. This setting ignores all but a small spot, 5 degrees or less in the center of the frame. The most common setting is center-weighted since it is less prone to either overly bright or overly dark backgrounds.

Some cameras have the ability to use the center-weighting concept but make it more flexible. The simpler of these allows the photographer to move the camera till the key subject is in the center of the frame and then press the shutter release half way. This holds both the focus (if auto focus was used) and the meter reading. Then the photographer can move the camera and compose the shot without the key subject having to be in the center. The focus and light settings will be optimized for the key subject. This is a very useful tool and it is easy to use. More complex solutions sense where the photographer's eye is focused as the photographer looks through the viewfinder and captures the light and focus information at that point.

Many separate, hand-held light meters can be configured for either incident light or reflected light. If the meter can read reflected light, it will have some sort of lens mechanism so that the photographer can be assured that the right subject item is being read. Incident light is read with a translucent dome over the sensor so as to capture all the incoming light, regardless of the angle of incidence.

Some light meters will read in lux (and some in the predecessor to lux, foot candles) and others will read in exposure values. The actual reading in the latter case is most often in terms of camera settings. The photographer dials in the sensitivity of the sensor, the ISO, and then he or she can read out the required shutter time for a given f/stop, or the required f/stop for a given shutter time. It might also read exposure value or, *EV*.

EV can be defined in terms of the amount of light that the camera will allow to fall on the film or sensor chip, or it can be defined in terms of the sensitivity of the sensor and the light available from the scene.

In the case of the light being allowed into the camera, the formula is (where $\log_2$ is the logarithm to the base 2):

$$EV = \log_2(f/^2) - \log_2(s)$$

where *f*/ is the *f*/stop of the lens, and s is the shutter time in seconds. This relationship can also be stated as:

$$2^{EV} = \frac{f/^2}{s}$$

In the case of the sensor sensitivity and brightness of the light, the formula is:

$$EV = \log_2(0.32 \times ISO) + \log_2(0.2918 \times cd/m^2)$$

where *ISO* is the sensitivity of the film or sensor chip and cd/m^2 is the level of the light in candela per square meter. If the light level is measured in lux, the second parenthetical term should be changed to (0.0929 lux). This relationship can be restated as:

$$2^{EV} = 0.0934 \times ISO \times cd/m^2$$

To make a proper exposure, the EV values from the two methods must be the same, which is to say:

$$\frac{f/^2}{s} = 0.0934 \times ISO \times cd/m^2$$

Rewriting this, we have,

$$cd/m^2 = \frac{10.71 \times f/^2}{s \times ISO}$$

This last equation makes it clear that, given the light available, the photographer can adjust the *f*/stop, or the shutter speed, and, in a digital camera, the *ISO* setting on the camera. The *f*/stop numbers are really based on a square root series, so increasing the value by one stop would reduce the amount of light coming into the camera by a factor of two. This could compensate for an increase in the light level by a factor of two. Alternatively, the shutter speed (in seconds) could be reduced by a factor of two. Or, again, alternatively, the *ISO* could be decreased by a factor of two. In addition, *f*/stop changes can be used simply to offset either shutter speed changes or *ISO* changes or some combination of both. This equation shows the basic relationships between the camera controls on the right of the equals sign and the available light on the left. Although lux is a measure of incident light and cd/m^2 refers to reflected light, the two are often compared. When this happens, it is instructive to know that 1.0 lux is equal to 3.1421 cd/m^2.

Figure 4.2 shows how *EV* changes as a function of lux for different ISO settings. Note the logarithmic relationship. Lux is linear and EV is logarithmic, like the *f*/stop series.

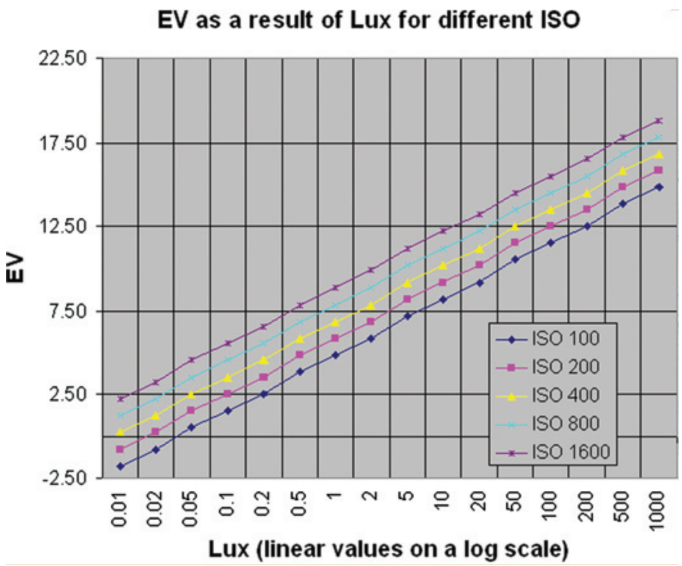


FIGURE 4.2 *Relationship between Lux and Exposure Value. The logarithmic relationship between lux and EV is shown for various ISO settings.*

TIME EFFECTS

In daylight exposures, the light remains relatively constant over time, and so the meter needs to give only a steady state reading. When an electronic flash is used, the meter needs to respond very quickly and integrate the light over time to give the total amount of light in just a few thousandths of a second. In many of the newer cameras the TTL meter can do this. In fact, when a camera-compatible flash attachment is used, the camera's meter will measure the light and when it reaches the right amount required for the given settings, a signal will be sent to the flash to quench the light and terminate the exposure.

When the automatic camera is used in the TTL mode, the photographer will have to select whether to fix the *f*/stop and let the camera select the shutter time, or fix the shutter time and let the camera select the *f*/stop. Another option is to use a program mode in which the camera automatically selects both the shutter speed and the *f*/stop in order to avoid unnecessarily extreme settings.

SPECTRAL COMPONENTS

If a light meter is calibrated to read in lux, then it will have filters over its sensor so as to provide an overall spectral sensitivity (sensitivity to each of the wavelengths) that is very similar to the human visual system. The human visual system has a slightly different response in bright light than it does in subdued light. The daylight response is called the *photopic* visual response and the nighttime version is the *scotopic* response. In photography, the photopic response is the one that normally is used.

If the metering is internal to a camera, then the sensitivity of the light meter will be designed to allow the camera to replicate the human visual system. Photographic films are supplied with separate layers designed to replicate the human visual system, and so metering should do the same.

The sensors used in both digital cameras and video cameras have native sensitivity in the infrared that is beyond the human visual system. Often an infrared blocking filter is inserted to prevent exposure due to this long wavelength light. In some video cameras, the filter is on a mechanical arm and can be taken out of the beam if the light levels fall too low. This is called a day/night camera. In reality it is taking nighttime images in the infrared, and not the visible portion of the spectrum. If a normal, lux-reading light meter were held in front of the subject when a day/night camera is working in night-time mode, it would probably read nearly zero lux. Cameras that use the infrared to take pictures are often referred to as "zero lux" cameras. This does not mean that the cameras do not need any light to take pictures; it simply means that they do not need any light in the visual portion of the spectrum, but they do need light in the infrared portion of the spectrum.

SPECTRAL SENSITIVITY

All light sources have different amounts of intensity (photons per square meter per second) at different wavelengths. All photo sensors have different amounts of responsiveness at different wavelengths, and the subject items have components that reflect different percentages of the photons in different portions of the spectrum. Finally, there is the possibility of inserting filters into the beam, which transmit different proportions of incoming intensity differently at different wavelengths. All these effects must be taken into account when determining the response that you will get from a given photographic setup.

The basic process is to start out by selecting a series of very narrow wavelength bands all the way across the spectrum. Then for the first band, determine the amount of light from the source, the light reflecting from the object, the amount passing through any filters, and finally, the response of the sensor. Now multiply all those values together. Next, repeat the process for the next band. Take the two products and add them together to get the total response for those two bands. The process is then repeated for each of the other bands all the way across the spectrum. The final sum of all the products is the response that one will get from that system. In principle, the bands can be made extremely small and the number of bands can be made very large. The result is that we can determine the overall response, R according to the following formula:

$$R = \int_{\lambda} (a \times Light) \times (b \times Object) \times (c \times Filter) \times (d \times Sensor)\, d\lambda$$

where R is the response and the light, image, filter, and sensor values vary with wavelength, λ. The coefficients *a*, *b*, *c*, and *d* are constants that depend upon the units of measure employed. Although the formula is rarely evaluated by photographers (it is used primarily by engineers designing photographic equipment and supplies), it is still instructive. The concept is important since it explains how a camera will respond when different light sources and filters are used with the same subject. It also helps show how different cameras or films might perform.

Consider the following photographic setup:

- A sensor with increasing response at longer wavelengths (like the silicon used in most digital cameras)
- A filter that passes lots of red light
- A subject that reflects red light
- A tungsten filament light source, which produces a lot of red light

The result will be a very strong response. If the light source were changed to mid-day northern daylight, which has more blue then red, the response will drop. If the filter were replaced by a green filter, the response would drop even more. The response adds together all the interactions at each of the wavelengths. The result will be responsive to changes in the spectral distributions of the components.

Chapter 5

Setting Exposures

THE SHOWER ANALOGY REMINDER

In Chapter 4 the analogy of the shower head was introduced. In this model, drops of water spray out from a source, getting farther and farther away from each other as they drop down. Gravity bends their flow paths downward, making the paths curve.

With light this does not occur. The photons simply follow straight paths, moving apart from each other as they move away from the source. With a point source, they radiate in all directions, creating a spherical wave front. The surface of a sphere is equal to $4 * p * r^2$, where r is the radius of the sphere. In the case of a steady light source, the number of photons per second coming from the lighting element is a fixed number. At distances close to the source, the number of photons per second per unit of area is a relatively high number. But if a point is chosen that is farther away, the same number of photons per second is now spread over a larger area, so the number per second per unit of area is comparatively lower. As the distance from the source increases, the area will drop off with the square of that distance compared to the previous distance because of the equation for the area of a sphere of a given radius cited earlier. The relationship is called the *inverse square law*. It holds when the size of the source is small compared to the distance from the source. In a typical room, it will hold for incandescent light bulbs, but not as well for fluorescent bulbs. It will not hold for focused beams like headlamps of cars or laser pointers. Light reflecting off a wall or ceiling will also not obey the inverse square law. With this key exception, the shower head analogy works pretty well and is a useful model to keep in mind when setting exposures.

A THOUGHT EXPERIMENT

Let's say that I go into a small theater, wearing sunglasses, and I take a seat in a choice spot about half-way up the rows of seats. The theater is empty so I can take my time and experiment. Only a few lights are on, so it is a bit dark, but not totally dark. On the stage is a series of gray squares hanging on the back curtain. Some of the squares are quite light in color and others are somewhat darker. There are six that are, for all intents and purposes, black. They all look the same and they are all black.

Now I take off the sunglasses and review the stage again. All the squares got lighter, and in fact now only four of the squares appear black. They are not very bright at all, but they are not black. When I turn up the lights, all the squares become gray. The ones that were black are all now a darkish gray, but none are black.

I go up on the stage and notice that each gray square has a density reading on its back. The darkest one is marked 2.3 neutral density. Another one is marked 2.15 neutral density, the next 2.0, 1.85, 1.70, and so on, in increments of 0.15 density units. Since density is a logarithmic scale, the equal increments tell me that the squares represent a series in which each successive square, going down the scale, is reflecting an equal percentage more light than the preceding. In fact each second successive square reflects twice as much light as its predecessor (e.g., 2.0 reflects twice as much light as 2.3, and 1.7 reflects twice as much light as 2.0).

So if the squares really provide incremental steps of reflection, why did I originally see that six of the squares seemed black, and then when I took off the sunglasses, two more turned from black to gray? The answer is that at the low light level originally present in the auditorium, and with the sunglasses on, the six darkest squares did not reflect enough light for my eyes to register a response. So they looked black. When I took the sunglasses off, the increased light allowed to enter my eyes apparently increased the light levels from the next two darkest patches to the point where they now registered a visual response. When I turned up the lights, all reflected enough light for me to see that they represented an ordered series of darkness.

Taking the thought experiment a bit further, assume that I light the stage with some very intense spotlights. Now when I look at the patches, three of them appear to be nearly pure white and the others have increasing levels of gray. In this situation, the bright lighting saturated my eyes. Consider the phenomenon called snow blindness.

There are two sets of factors being addressed in this scenario. First there is the comparison between the sensitivity of my eyes and the amount of light available on the scene. Second, there is the exposure setting (in this case the imposition of the sunglasses). The same factors must be considered when setting a camera: the light on the scene compared to the sensitivity of the film or digital camera, and the lens opening in combination with exposure time.

In this chapter we will pull together the notions of dynamic range, exposure setting, shutter speed, and certain lens effects.

TWO VIEWS OF EXPOSURE

In Chapter 4 it was noted that there are two ways to evaluate exposure. One way to measure exposure is to take account of the amount of light that the camera will allow to reach the sensor—that is, the size of the aperture and the time during which the aperture is open. This is analogous to the size of the opening of the can in the shower and the time during which there is no cover over it. This measure does not explicitly take into account the sensitivity of the sensor in the camera or the brightness of the scene.

The other way to measure exposure is in terms of the sensitivity of the sensor and the brightness of the scene. In this situation, we want the range of scene brightness to be such that, for the sensitivity of the sensor, a suitable image will be obtained. This means that the range of brightness levels of the important key objects in the scene fall on the monotonic range of sensitivity of the sensor, meaning, within its dynamic range. Failure to accomplish this was described in the theater thought experiment. When setting the exposure parameters of the camera, we are ensuring that the range of brightness levels in the scene, as modified by the camera's aperture and shutter time, is such that it is consistent with the dynamic range of the sensor.

In Chapter 4 it was shown that this occurs when the exposure value (EV_a) based on the amount of light being allowed into the camera is equal to the EV_b we obtain when considering the sensitivity of the sensor (ISO) and the brightness of the light available. The two determinations of EV are:

$$2^{EV_a} = \frac{f/^2}{s}$$

$$2^{EV_b} = 0.0934 \times ISO \times cd/m^2$$

Then, setting them equal to each other gives us:

$$\frac{f/^2}{s} = 0.0934 \times ISO \times cd/m^2$$

which can be rewritten as,

$$cd/m^2 = \frac{10.71 \times f/^2}{s \times ISO}$$

which we will call the exposure setting equation.

To understand the exposure value derived from the scene brightness and the sensor sensitivity, EV_b, we need to refer back to the thought experiment. Areas of the scene that are very dark will not be rendered by the sensor if they do not present enough light to the sensor at its low-threshold sensitivity point. The scene brightness is indicated by the term on the left side of the equation. If there is not enough light, additional light must be found, either by studio lights or a flash attachment of some kind.

The right side of the equation deals with the question of how much light is enough. If the f/stop is set to its lowest value (as open as it can get), the shutter speed is set to its longest open time, and the ISO (on a digital camera) is set to its highest value, and the equation still does not balance, there is not enough light. Consequently the same thinking can be applied in reverse if the light is too bright.

At the same time, the bright portions of the scene should not present light levels that are beyond the saturation point of the sensor, and the dark levels cannot be below the threshold level of the sensor. In other words, the scene's range of brightness levels should be such that they fit within the dynamic range of the sensor. If the scene's brightness range is less than the dynamic range of the sensor, there is no problem. Adjustments can be made when the image is processed. If it is greater, there is a problem, since certain scene information will never be captured in the first place, so no degree of postprocessing of a single image can compensate.

It is best to reconsider the situation of taking a wedding photo of a bride and groom. She is wearing a white-on-white dress that is extremely important to her, and he is wearing a black-on-black tuxedo of which he is very proud. You need to take a photo that will make them both happy. Now, if the scene's range of brightness is greater than the dynamic range of the sensor, the photographer will have to make a choice from the following short list:

1. Lose detail in the dark areas in order to preserve detail in the bright areas (tuxedo is all black but the dress detail is preserved so she is happy and he is not).

2. Lose detail in the bright areas in order to preserve detail in the dark areas (dress is all white but the black velvet lapels are distinct from the jacket body so he is happy but she is not).

3. Compromise and lose a little of both the bright and dark areas (probably neither is happy).

4. Take two shots, one per the first scenario and the other per the second and merge the two at a lowered overall contrast ratio. (Will they stand absolutely still for two consecutive shots?)

There are not very many proper solutions. The professional wedding photographer generally will use a "portrait" film that has an extended dynamic range, or if he or she is using a digital camera (digital cameras have significantly less dynamic range than you can get with portrait film in a film camera), will position additional (fill) lights shining on the groom. This will make the groom lighter compared to the bride, and thereby decrease the scene's range of brightness levels.

Whatever the practical solution, the idea is to line up the scene's brightness range with the sensor's dynamic range. This is done by adjusting the aperture, exposure time, and ISO (after the additional fill light is brought in). Many modern cameras will make the adjustments for you if you use the auto-exposure mode. Unfortunately, cameras are not sensitive to the feelings of the bride and groom, so some degree of manual intervention may be needed.

SOME COMPLICATIONS

How do we determine scene brightness? In the shower analogy, the number of drops per unit of area per second was suggested. In this analogy, the photometric unit of measure would be lux. But this is the light falling on the scene. In setting exposure, we need to know the light leaving the scene and going to the camera. Extending the shower analogy, assume that the drops coming from the shower head passed through a very cold room on the way to the can, which is placed upside down. The result is that the drops freeze into solid pellets, forming hail, on the way to the can, and when they get to the floor of the shower, they bounce off the bottom of the can (the bottom is facing upward). The rate of the bouncing hail is the surface brightness and it is measured in foot-Lamberts or candela/square meter, as if it were a source of light. If the surface were 100% reflective, the number of foot-Lamberts would be given by 0.0929 times the number of incident lux. So the earlier equation:

$$EV_b = Log_2\left(b * 0.32 * ISO\right)$$

can be written:

$$EV_b = Log_2\left(b_L * (0.32 * 0.0929) * ISO\right)$$

where b_L is the scene brightness in lux.

Another complication is that, as written, the camera setting process is based upon a particular point on the brightness scale that corresponds to a particular location in the scene. Instead, we need a reading that is more indicative of the average. To deal with this, we use an 18% gray card. Eons ago it was determined that the world is, on average, 18% gray. That is, it reflects 18% of the light landing on it over a goodly range of the visible spectrum. So if you are making incident light readings in lux, you will need to estimate that only 18% of that light is really available to the camera. If you can use a spot meter that measures the reflected light from the scene at various points on the scene and calculate an average reading, that will be more accurate and it will probably be in cd/m^2. The automatic cameras that read the light with TTL meters do this automatically. With a digital camera, we can let the camera set itself automatically and then take additional pictures at higher and lower EV_a levels and pick the best one. This process is known as an *exposure series* or *bracketing*. If the subject is moving, however, this could be a serious problem.

EXPOSURE TIME EFFECTS

There are some issues related to exposure times that are more than just setting the proper level of exposure. As a general rule, with modern cameras, the shorter the exposure time, the better. Consider the other extreme. You are doing the wedding picture noted earlier and arrange for a six-minute exposure. After just a few seconds the pair starts to fidget and after several seconds, they start to complain. Finally, after a few minutes they walk away. What have you got in the picture? Nothing but a blurry mess. The picture the camera takes is a compilation of everything that went on while the lens was open (and the lights were on). The shorter the time, the less movement could have occurred. One common problem is to have the flash on and have a portion of the scene brightly illuminated by external light. The exposure time might be somewhat longer than that of the flash duration, but if any movement occurs, the image will have blurs. Typically, electronic flash devices send out the light for only a few thousandths of a second. They are very bright and good at stopping action. But if you manually set the exposure too long, you could have problems. With the better electronic flash devices used with TTL cameras, the reading of the light inside the camera actually turns off (quenches) the flash once enough light has been achieved.

As is clear from the exposure setting equation, the shortness of the exposure will be dictated by the choice of f/stop, the amount of light available, and the sensitivity of the sensor. It might also be dictated by the choice of lens. Telephoto lenses magnify the effect of movement of the camera. So if the camera is not mounted on a steady tripod or other such device, a shorter shutter time will be needed. The rule of thumb is to take the focal length of the lens in millimeters and divide it into one. This gives the longest shutter speed you should use. For example, if using a 300 mm lens, you should have a shutter speed of 1/300 seconds or less; with a 600 mm lens, 1/600 seconds, and so on.

Another problem with long exposure times is increased noise. Digital cameras use sensors that create small electrical currents when exposed to light. Unfortunately, they also have a nasty habit of doing so in the dark as well, a so-called dark current. Modern camera sensors have very low dark currents compared to their normal operating range in the light, but if the exposure lasts for several minutes, the accumulated effect of the dark current flowing for the extended time can be meaningful. It will show up in dark areas of the scene and it will take the form of a multicolored speckle pattern and will result in lower-than-expected average darkness values. The multicolored salt-and-pepper pattern in dark portions of the image is due to the fact that the dark current effect varies from pixel to pixel. Any attempts to bring out detail in these areas will be hampered by this random noise pattern.

F/STOPS

Chapter 3 introduced f/stops; they are mentioned again here because of the many ways in which they affect the image, beyond just helping to achieve a proper exposure. Specifically the choice of the f/stop will affect the depth of field, the resolution, and, of course, the shutter speed.

The angle at which light rays cross is greatest when the lens aperture is at its largest setting (lowest f/stop). As a result, if the sensor is not quite at the point of that crossing, the image will blur. Conversely, when the lens is at a smaller opening, the angle is lower and the rate at which the image blurs with deviation from the point of ray crossings is lower. These effects translate into distances on the object side of the lens. What is changing is range of distances at which object items can be located and still be in sharp focus.

If the photographer wishes to take a photo that is concentrated on a relatively near object and not have distractions from background items, he or she can use a large aperture (low f/stop number) and focus on the near object. Conversely, if the photo is intended to show several items at varying distances, a small aperture (high f/stop number) should be used. In any case compromises will be necessary because of the amount of light, the sensitivity of the sensor, and the shutter speed requirements.

Theoretically, the resolution of an optical system goes up as the aperture increases. This is why astronomical telescopes have such large lenses. Telescopes are in reality giant digital cameras, and actually they usually employ mirrors instead of lenses, but the same principle applies. Having a large number of rays converge gives better resolving power than having fewer. It also captures more light (the can on the shower floor has a bigger opening). However, most glass lenses contain imperfections. If the number of imperfections is uniform per unit of area of lens, then larger apertures will be subject to more imperfections. According to this reasoning, a small aperture should be used. Obviously a compromise is in order. The most common rule of thumb is to use a setting that is one-half to one-third the way (in f/stop notations) between the largest and the smallest aperture for a given lens when other conditions allow.

ADDITIONAL LIGHT

By and large, the aperture and the shutter speed allow the photographer to reduce the amount of light getting into the camera. These are the tools of existing light. When existing light is insufficient to make a good exposure, light must be added. This can be done with both flash devices and steady sources such as incandescent lamps. In the case of steady sources, it may well turn out that a significant amount of power is required. First of all, the light is on all the time, and second, there is the logarithmic nature of the system. If the light is doubled by adding light from a 250 watt source, that will buy one f/stop. To get to two stops, two such lights are needed. To get to three stops, 1000 watts are needed, and so on. The next increment will blow the fuse in the typical household circuit. This sort of incremental lighting is difficult to achieve at a crime scene.

More commonly, a flash source is used. These can put out a very bright beam of light, but do so for only a few thousandths of a second. They run of batteries, but can add large amounts of additional exposure. These units are rated by their "guide number," which will be provided by the manufacturer. To obtain the guide number, take a series of photos using ISO 100 with different f/stops. The distance to the subject is recorded. Repeat this at a number of distances, and note the f/stop that gives the best images at each distance. The guide number, ng, is defined as the distance, d, times the f/number; that is:

$$N_G = d * f/\#$$

The average of the values obtained is what is published. The photographer can use the guide number to determine the f/stop given the distance, or at a given f/stop how far away the subject can be from the camera. That is,

$$d = N_G/f/\# \text{ when solving for distance, and}$$
$$f/\# = N_G/d \text{ when solving for f/stop}$$

Modern, automatic cameras with TTL light metering can deal with this without any intervention on the part of the photographer, but there are always special situations that require a little arithmetic.

PROCESS DIAGRAMS

To help visualize the exposure setting process, a graphical representation (often used in photography) is depicted in Figures 5.1 through 5.4. Figure 5.1 shows the basic arrangement of the process but with no exposure being made. In the upper left-hand corner is a graph that shows the basic sensitometric response of a hypothetical digital camera. The input is the scene, here shown as a series of rectangles of increasing brightness. The input to the camera's sensor is shown on the horizontal axis as Log Exposure (base 10). The output from the camera is shown on the vertical axis as Value. This is a digital value ranging from 0 to 255 for an 8-bit camera, and 0 to 4096 for a 12-bit camera. Higher levels of exposure result in higher output values.

In the upper right corner is the computer. This takes values as inputs and also outputs value. It has the ability to transform these numbers. That will be shown as a contrast shift later in this series of graphs. In all the graphs in this series, higher input values result in higher output values.

In the lower right is the printer. It takes values as inputs and outputs colorant deposited on paper. In the case of a printer, the lower the input value, the higher the amount of colorant on the paper.

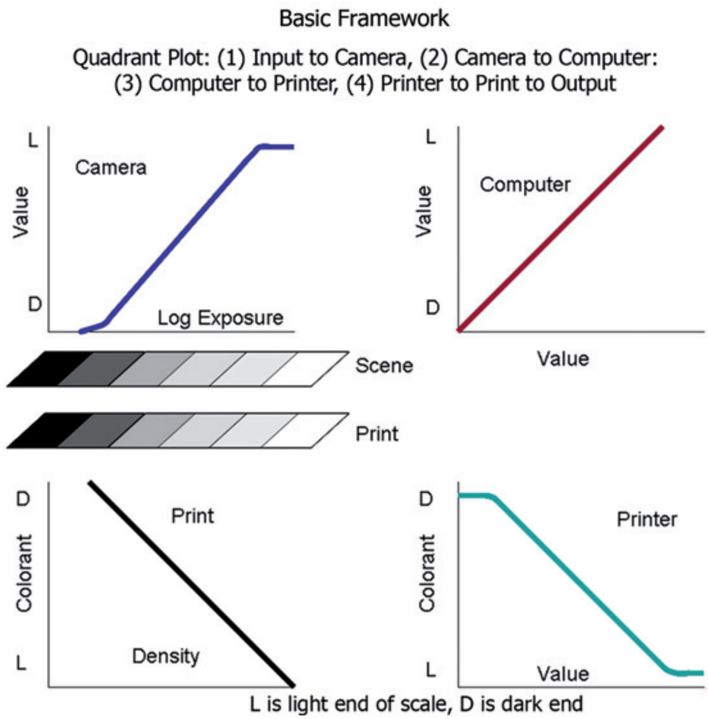


FIGURE 5.1 *Basic Tone Reproduction Framework. Four key elements—the camera, computer, printer, and print—are shown as they respond to the scene. The approach helps to visualize how the different elements influence the final outcome. Only the individual element responses are shown in this figure.*

The final graph shows the print itself. For the sake of convenience, it allows presentation of the printer output just below the original scene. It is derived from the original scene inputs and printer outputs. Note that the input to this function is on the vertical axis and the output is on the horizontal axis. In this rotated orientation it shows the full system output compared to the system input.

Figure 5.2 depicts the act of exposure. Light from the scene goes through the lens (not depicted) and strikes the sensor. This is depicted as vertical lines going from the patches to the sensor response curve. This results in a series of values, each resulting from one of the scene patches. The values, shown as horizontal lines, exit the camera and go to the computer. In the current case, the computer simply outputs the same values coming in. These are shown as vertical lines going out of the computer and into the printer. The printer responds to the input values by placing colorant on paper. The colorant levels are shown as horizontal lines going from the printer and to the print quadrant. The points where the print output levels cross the input scene levels are connected to create a system reproduction curve. Note that the output levels are on the horizontal axis. These point to the reproduced patches. Note

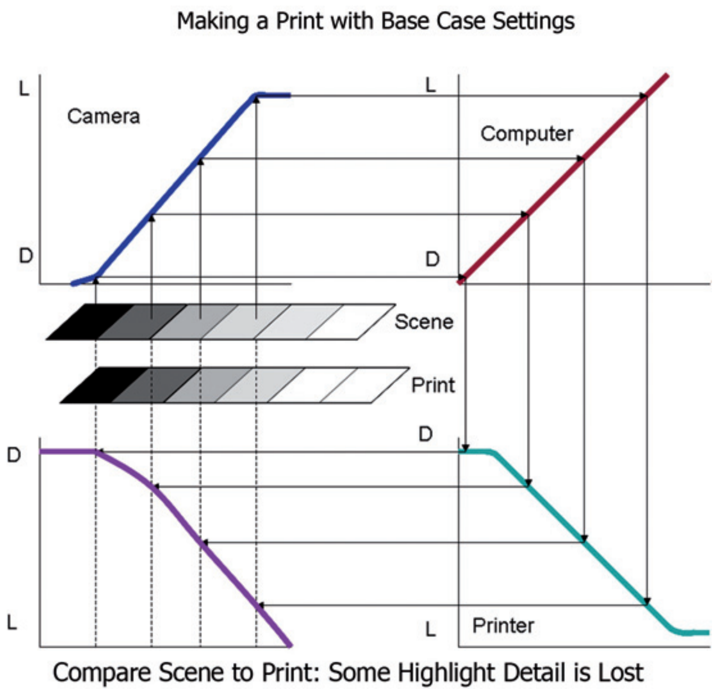


FIGURE 5.2 *Making a Print. Following the format of Figure 5.1, the effect of the flow from the original scene all the way through the system is shown. The arrows indicate the direction of the flow.*

that the original scene has a greater range of brightness levels than the system can reproduce. The result for the current settings is that the lightest patch is reproduced at the same level as the second lightest patch. Some information has been lost. Had the lens been set to a smaller aperture, one that is equivalent to one patch level, a set full of patches would effectively have been slid over to the left, and the output would show that both of the two lowest patches are black and the brightest two patches would have different brightness levels. Again, some information would be lost, but it would be shadow information instead of highlight information.

In Figure 5.3 the computer is programmed to give a proportionately larger output relative to its input. This is an increase in contrast. Again the results for each patch within the camera's dynamic range flow through the series of charts to the output. The end result is that the system reproduces patches from black to white in fewer steps, which is to say that the contrast has been increased.

In Figure 5.4, the original exposure level and the normal contrast level are restored, but the computer is programmed to add a constant to each incoming value. Graphically, this shows up as a shift of the computer's response line to the right. Following the tracings on through, it is clear that the patches that are reproduced are rendered with higher brightness levels.

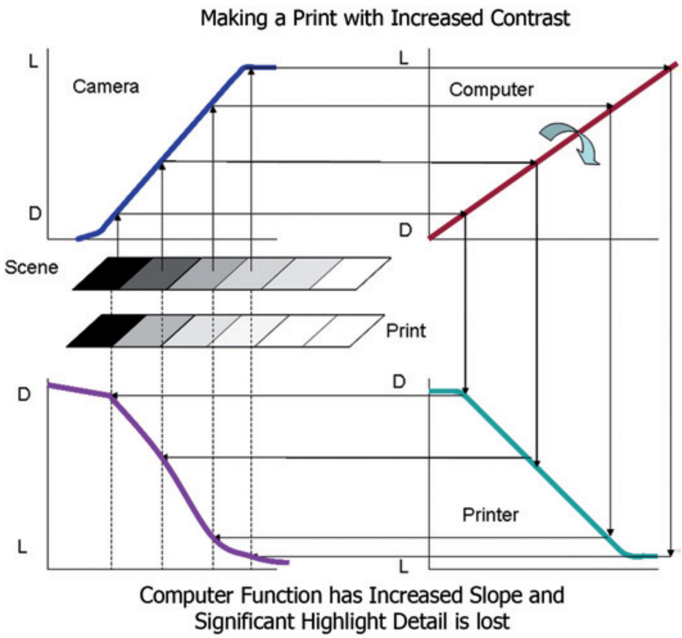


FIGURE 5.3 *Increased Contrast. The computer's response function is changed, resulting in more output per unit input. Ultimately this produces larger increments in the steps in the print and, in this case, a loss of highlight information.*

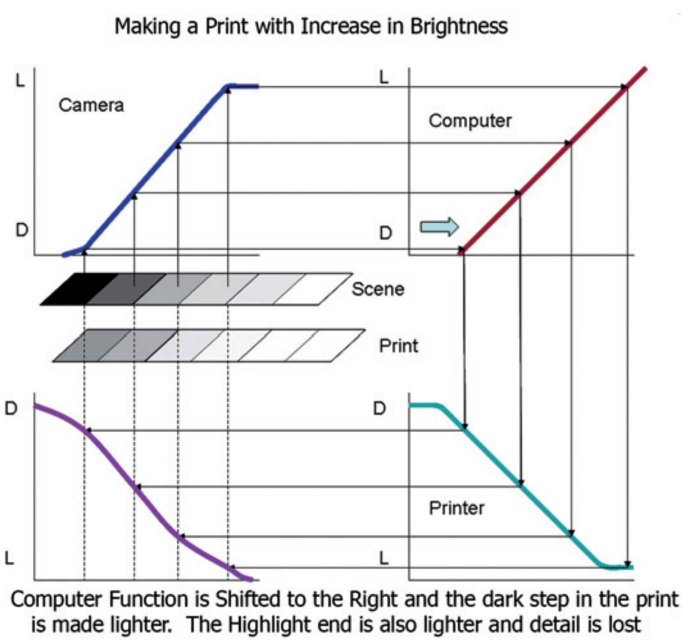


FIGURE 5.4 *Increased Brightness. The computer's response function is shifted to the right, adding a constant to each incoming value. The ultimate result is a lightening of all the steps, with, in this case, losses in the highlight information.*

This is consistent with the notion that adding a constant results in a brightness increase.

These figures are here to help you visualize what happens as exposure is changed, how the dynamic range of the sensor interacts with the brightness range of the original scene, and how the computer can be used to modify the outputs with given inputs. Note that the computer cannot bring back information that was lost because the dynamic range of the camera was less than the brightness range of the original scene.

SETTING UP EXPOSURE PREFERENCES

Most modern digital cameras have at least four settings for how the automatic exposure system will operate. The photographer must select one.

1. The most common is Program Mode. In this mode, the camera will measure the light and select nominal aperture and shutter speeds to get a good image. Rules of thumb will be followed as the light level permits. The photographer does not have to make any further settings. This will avoid extreme values in both shutter speed and aperture.
2. When the subject is moving it will be necessary to use a high shutter speed: a small fraction of a second. Since this is a requirement, the aperture must be set to compensate. For this, the Shutter Preferred Mode or Shutter Priority Mode is used. In this mode the photographer selects a shutter speed and the camera sets the aperture. If there is a lighting limitation, it can result in very wide apertures.
3. When the photographer wants to have either a large depth of field or a small one, he or she will have to set the aperture accordingly. In these situations, the Aperture Preferred Mode is selected. The photographer selects the aperture and the camera sets the shutter speed accordingly. Again, depending on the lighting, the camera may select a very short or very long exposure to compensate for the selected aperture.
4. To deal with other, more unusual situations, the photographer will choose the Manual Mode. In this case the photographer has to select both the shutter and the aperture settings.

Highly automated cameras must take into account the ISO setting put into the camera by the photographer and the lens focal length (read automatically when a lens is mounted). When lens extenders are used, the automatic controls normally do not work and the camera will have to be in manual mode.

When a scene is highly variable, special care is necessary. The camera's metering system typically can work in three different modes. One averages the entire frame (*full-frame*), another reads the full frame but gives more weight to the material near the center (*center-weighted*) of the frame with the tacit assumption that this is where the key subject is to be found), and the third reads only a small spot in the center of the frame (*spot-metering*).

When taking a photo of two people standing next to each other where there is a dark background behind them, a special technique is necessary. If spot-metering is used, the camera will see only the dark background and will greatly overexpose the two people. If it is set to center-weighting, the results will be better, but still not optimal. Overall averaging is often subject to a large number of nonimportant features in the image, so it is not a good choice either. To deal with this rather common situation, the cameras usually have a shutter release with a partial-pressure position. The photographer can select center-weighting, point the camera at one of the people in the image, and press the release halfway down. Then, while holding the button halfway down, move the camera to the desired composition and press the shutter release the rest of the way. The camera will hold both the focus and the exposure settings obtained when halfway down, and then apply them for the final release. Once the picture is taken, the settings are discarded and the next image is ready for a fresh start.

QUESTIONS TO CONSIDER

1 Assume it has been decided that an optimal exposure of an object 6 feet away is f/2.0 and 1/200th of a second. However, it is then necessary to move 12 feet away to actually take the photo. The shutter speed of 1/200 needs to be retained since a 200 mm lens is being used. Does the f/stop need to be changed, and if so, to what value?

2 A zoom lens is being used, and its aperture is set to wide open at f/2.0. Originally it is at 50 mm. But it is then adjusted to 100 mm. Will the lens have the same f/stop, or will it be different? If different, what will the new value be?

3 At a crime scene, there is a navy blue hat, a bit of chain, a small pocket knife, and blood-stained black gloves on a white bed sheet. Should the photographer use auto exposure? Why or why not? If auto exposure is used, what might the image be like?

4 At a crime scene, there is a white-on-white handkerchief, a business card, and a pen knife lying inside a black-lined suitcase. Should the photographer use auto exposure? Why or why not? If auto exposure is used, what might the image be like?

5 At an outdoor homicide scene, there is some clothing stuck up in a maple tree. A 300 mm lens will be used to get a close-up of the clothing. What

is the longest shutter speed that should be used? If it is summertime, can auto exposure be used? What about if it is winter time?

6 Let's say you received 100,000 general photos from as many different outdoor crime scenes and averaged them all together. What would the resulting image be like?

7 You are at a scene and have a spot photometer. You make some measurements and find that the darkest items that you want to capture in the scene measure 4 foot-Lamberts. The brightest items you want to record in the same image measure 12,000. Your camera has a dynamic range of 2500:1. The camera is on a tripod and nothing in the scene is moving. How will you take the picture, and why use that process?

EXERCISES

Depth-of-Field Exercises

Depth-of-Field is controlled by three factors:

1 Focal length of the lens. For any given distance and aperture setting, the wider the focal length of the lens the more depth-of-field.
2 Distance from the subject. For any given focal length and aperture setting, the more distance between the camera and the subject the more depth-of-field.
3 Aperture setting (f/stop). For any given focal length and distance, the smaller the aperture the more depth-of-field.

Exercise 1: Outdoors

1 Place objects at varying distances from the camera (approximately 8 feet to 100 feet).
2 Use the normal angle lens for your camera.
3 Place camera on a tripod.
4 Compose the frame so all objects appear in the viewfinder.
5 Set the camera to aperture priority.
6 Set the camera to manual focus.
7 Focus on the object at approximately ⅓ the distance into the frame (30 feet).
8 Make photographs at: f/3.5—f/5.6—f/8—f/11—f/16—f/22.

Exercises 2 and 3: Outdoors

Repeat the same procedure as in Exercise 1 for these two exercises:

1 Using the wide-angle lens.
2 Using the telephoto lens.

Note: *The primary control is the aperture setting*. Increasing distance decreases image size. Wide angle lenses distort the subject and telephoto lenses compress the image.

Stopping Motion Exercises

Stopping motion is controlled by the shutter speed of the camera. How fast a shutter speed is needed is controlled by three factors:

1 How fast is the action?
   - The faster the action, the faster the shutter speed necessary to stop it.
2 What is the distance from the camera to the action?
   - The closer the action is to the camera, the faster the shutter speed necessary to stop it.
3 What direction is the action traveling in relation to the camera?
   - Action traveling parallel with the camera will require a faster shutter speed.
   - Action traveling toward or away from the camera will require a slower shutter speed.

- Action traveling diagonally to the camera will require a shutter speed in between the two previous examples.

Exercise 1: Outdoors

1. Find a safe area along side a busy street.
2. Use the normal angle lens on your camera.
3. Place camera on a tripod approximately 40 feet from the street.
4. Ensure the camera is parallel with the street.
5. Compose the frame on the middle of the street.
6. Set the camera to manual focus—prefocus on the middle of the street.
7. Set the camera to shutter priority.
8. Make a photograph at shutter speeds of 1/30—1/60—1/125—1/250—1/500—1/1000.

Exercises 2 and 3: Outdoors

Repeat the same procedure as in Exercise 1 for these two exercises:

1. Stand alongside the street with the motion moving diagonally toward the camera.
   - Use the normal angle lens.
   - Place camera on a tripod.
   - Compose the frame—prefocus the camera.
   - Set the camera to shutter priority.
   - Make a photograph at shutter speeds of 1/30—1/60—1/125—1/250—1/500—1/1000.
2. Stand alongside the street with the motion moving diagonally away from the camera.
   - Use the normal angle lens.
   - Place camera on a tripod.
   - Compose the frame—prefocus the camera.
   - Set the camera to shutter priority.
   - Make a photograph at shutter speeds of 1/30—1/60—1/125—1/250—1/500—1/1000.

CHAPTER 6

Resolution

Silver halide film is composed of a thin layer of gelatin coated on a transparent film base. The silver halide is in a crystalline state and the crystals range from small to very small, finer than powdered sugar crystals, and only a few microns across. They are dispersed several deep in a layer that is a few microns thick. The crystals are activated when struck by a photon; once activated, they can easily be converted to metallic silver by a developing agent. The resulting microscopic grains of silver appear black in the processed film. In color films, a dye is formed by developer as it converts the silver halide to silver. Afterward, the silver is extracted and dye globules are left in the film. Since the image is formed by globules that are just above molecular size, these films are capable of rendering very fine details.

Digital cameras rely on a sensor chip. This is similar to a computer chip on which light-sensitive patches are formed in a regular array, usually in rows and columns. The set of patches is similar to a miniscule tile floor. Each patch is a light meter. It reacts to incoming light by creating electrical charge. The amount of charge is then measured for each patch and a number proportional to the charge is recorded. The ordered set of numbers is the digital image file.

Digital images comprise a rectangular array of very small patches, called picture elements, or pixels. Each pixel in the final image holds the color information for its location. The result is essentially a mosaic: each tile (pixel) holds only one color (composed of specific amounts of red, green and blue). And it is the combination of these that comprises the image that viewers see. There is no detail in a digital image that is finer than a single pixel. The ability of a sensor chip to render fine detail depends upon the number of pixels it has.

BASICS OF RESOLUTION

In Chapter 3, the idea of resolution was introduced briefly. In this chapter we will develop the means to calculate the resolution we can get with a given

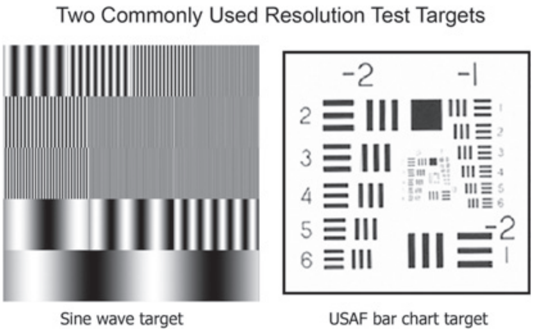


FIGURE 6.1 *Two Commonly Used Test Targets. There are many test targets in regular use. These two are common and very different. The sine wave pattern is used to measure the Modulation Transfer Function and the three-bar USAF pattern is used to measure line pairs per millimeter that are resolved.*

camera. In the case of film cameras, the resolution can be changed by the choice of film. But in digital cameras the sensor is a physical part of the camera and so is not adjustable. (Many digital cameras allow the photographer to save images at a lower resolution than the full capability of the sensor chip, but not higher.) Another important factor is the lens quality. For most of this chapter, it is assumed that the quality of the lens is so good that it is not really a factor. But beware, some aftermarket lenses are fairly poor and can have an impact on the resolution you can achieve with a given camera.

The place to start is with some basic phenomena and the terms associated with them. First there is the issue of spatial frequency. To understand this, assume a repeating pattern such as the bar charts or sine wave charts in Figure 6.1, which show two common implementations.

Basically, as the feature size gets smaller, the spatial frequency gets higher. In the case of sine waves, the wavelength λ is the distance from one peak to the next. The frequency ν is the reciprocal of the wavelength, as follows:

$$\lambda = 1/\nu$$

The wavelength is in units of distance such as inches or millimeters, and the spatial frequency is in reciprocal distance such as per inch or per millimeter. Low frequency information has a long wavelength and corresponds to the broad portions of the image. High frequency information has a short wavelength and corresponds to the fine markings in the image.

In the test targets, selected frequencies are presented and the results in the photograph are compared with the original target, frequency by frequency. In general, as the frequency increases, there will be an even, high level of ability to reproduce the target up to some point. Beyond this point, the ability of the system to reproduce the target will decrease.

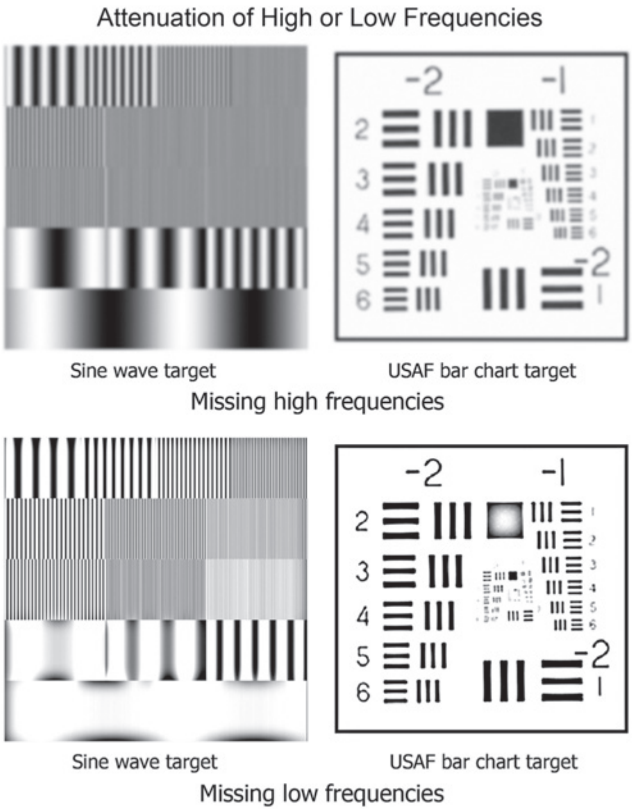


FIGURE 6.2 *Attenuation of High or Low Frequencies. The upper rendition shows the loss of high frequency information and how it makes edges soft and fine lines disappear. The lower rendition shows the loss of low frequency information and how it hollows out solid areas and makes broad sine wave bars disappear.*

The tendency is to think in terms of the rows and columns of the pixels. But the real world is more complicated than that. The object might be composed of lines, but they are probably not aligned with either the rows or the columns. The processes involved work about the same in both the rows and columns (except in certain video applications where the pixels are not square). So if the angle is known, we can estimate the resolution with Pythagoras' Theorem. This implies that the best result we might hope for is that the lines are parallel to either the rows or columns since the hypotenuse is greater than either of the other two sides of a right triangle.

Note the two images in Figure 6.2: the top one is missing high frequency information and the bottom one is missing low frequency information. The original did not have patches with discrete frequencies, however. What the figure demonstrates is that the various frequencies behave as though they were built into all the features in the image, and that they add together in special ways to form more complex structures. For example, sharp edges of even broad elements such as the block letters are built up of a large number of high frequencies all added together (Fourier's Theorem). When the higher frequencies are selectively decreased or eliminated, the edges become blurry.

The higher frequencies are not presenting the broad, flat areas, however, and so these are unaffected. The broad flat areas are comprised primarily of lower frequencies, and when these are attenuated or blocked, the broad areas become hollowed out.

With traditional film photography, where the photographer can change films, the important factor (again assuming a very good lens) is the film. Accordingly the ability to reproduce detail is stated in terms of distances *on the film*. It might be stated as 100 line pairs per millimeter on the film. With digital cameras, this is not the case. There is no film that can be changed to give a higher resolution and we will never work directly with the image on the sensor chip, so such a rating is of no interest. Instead, we must be concerned with distance on the object. So the limit will be in terms of, let's say, line pairs per millimeter *on the object*. An alternative is to work in line pairs per frame width. Then assuming knowledge of the frame width in millimeters, for example, it is simple to compute the number of line pairs per millimeter on the object. The real question is, "If I take a photograph of a full shoe impression, will I be able to resolve the fine details and make a unique identification?" If the shoe is a small size and the details are of a goodly size, the odds are good. If the shoe is much larger and has the smaller detail, the odds will decease.

RESOLUTION CAPABILITY

The resolving power of a camera is measured by the size of the finest details that are separate entities in the object and can be represented as separate entities in the image. The modulation transfer function (MTF) does not give a set limit, but rather portrays the degree to which the camera represents sine wave objects in the image. Other related terms are acutance and acuity; these will not be discussed here.

As one records higher and higher frequencies, the ability of a sensor to reproduce detail decreases, and sooner or later it will reach a point where the detail is either not reproduced, or it is reproduced so poorly that we can consider the feature not reproduced. The point just before this complete failure typically is cited as indicative of the power of the sensor to resolve detail. The standards for determining the resolving power cut-off point vary depending upon the target being used, the criterion the investigator chooses to use, and the way in which the sensor performance deteriorates. They are:

Loss of modulation. This typically is used when measuring sine wave responses. The modulation of the original is the peak density of the chart minus the trough modulation of the chart. Similar measurements are made on the image and the ratio of the two is given, usually in log base 10 values (audio devices are measured in decibels, which is also a log base 10 measure). When the log loss in modulation is equal to 0.3, the

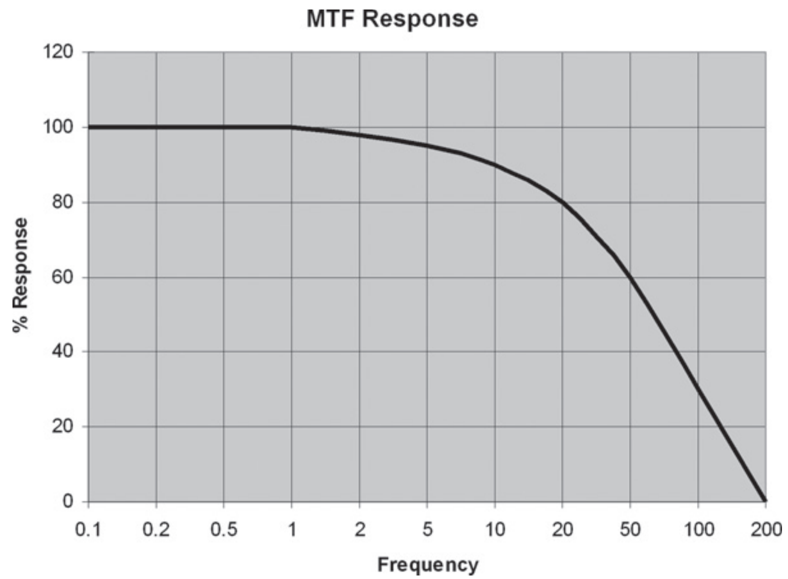


FIGURE 6.3 *MTF Response. In unfiltered images, there is a tendency for low frequency information to be reproduced at 100%. For higher frequencies, the response rate drops off. MTF curves for the elements of an imaging chain can be multiplied together to give the overall response function of the system.*

actual modulation has dropped by half. Sometimes this is used as the cut-off point. A more generous analysis results if we compare the modulation due to the image to the modulation that is the result of random noise. When the noise level is 0.3 log units lower than the target-induced modulation, then the signal-to-noise ratio is 2:1, and this might be used as a cut-off point as well.

When using this approach, we speak of the modulation transfer function, or MTF of the "system." To obtain this, we measure the log modulation compared to that of the target at several frequencies. The result is a curve that is flat up to a point and then starts to slope downward. Figure 6.3 shows a MTF curve. As it turns out, the MTF for a full system is the product of the MTFs of the elements: the sensor, the lens, the printer, and whatever else is in the system. On a log scale, the values for all the elements can be added and then normalized to get the overall response since this is the equivalent of the product of the values in linear space for each of the elements. The result is that we can measure the various components of the system and compute the system MTF, or we can infer the MTF of one of the elements if that of the system and the other elements is known. In practice, MTF is useful for engineers designing elements of a system, but not very practical for the camera user.

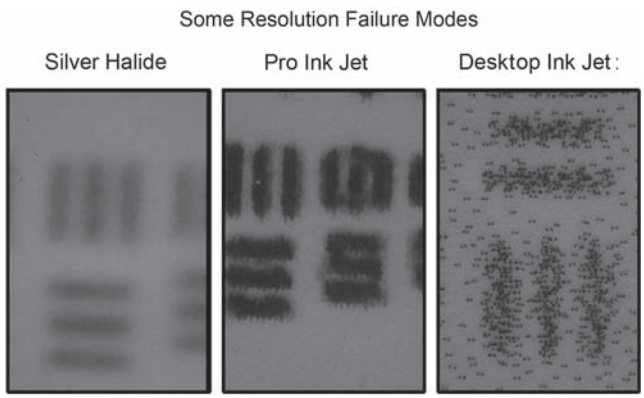


FIGURE 6.4 *Some Resolution Failure Modes. Photomicrographs of three-bar resolution test targets at points close to failure. The silver halide image shows typical MTF roll-off in the form of a blurring of edges. The Pro Ink Jet printer shows high density areas blocking and merging across the space between bars. The Desktop Ink Jet printer shows the generation of a lot of random noise, making it hard to know which spots go with which bars.*

Increase of Noise. In portions of an image with low levels of light, the signal generated by detail in the object will result in small levels of response from the sensor. It is also at these points where pixel-to-pixel variation due to random noise is most noticeable (due to dark current). The noise results in a colored version of salt-and-pepper patterning. At some point, the average level of the noise is comparable to that of the signal, at which point it is not possible to know whether a given feature in the image is due to signal or noise. Typically we look for a signal-to-noise level (averaged over a small area) of 2:1 or better. Whether we are looking at sine wave or bar chart test targets, the result is the same: we cannot determine what is being shown in the image.

If the image is on a print, then typically there is a problem at the high-brightness areas of the image. The noise level goes up in fine detail areas due to random fluctuations in the printing device and the level of modulation is down. The result is that we cannot determine whether the areas between the supposedly separate dark areas on the test target are being seen as separate or not. Figure 6.4 shows this effect. In any event the system has failed to clearly show the separate features in the input as separate in the output.

Pixelization. In digital images, the picture comprises millions of small picture elements, or pixels, each of which has a single color. If the image is enlarged multiple times on a computer screen, it is often possible to see the separate pixels. There is no detail smaller than a pixel. But any feature in an image has more than a single color. A yellow spot on a brown background might comprise a feature. Without a background of a different color, the feature is merely another spot among many that are all the same. Typically the smallest feature is portrayed as a dark line adjacent to a light line, or a line pair.

Clearly a number of line pairs can be lined up so that the dark and light lines alternate. The result is a grid of lines. The United States Air Force chart has such a pattern. There are three line pairs in each patch and there are pairs of such patches near each other but laid out so that they are perpendicular to each other. The patches vary systematically in size, and we look for the smallest pair of patches where both sets of lines are all separate and distinct. The size that corresponds to this set of patches is what is used for stating the resolution of the system under test.

Combinations of criteria. A typical combination of problems might occur in photographing a shoe impression. The fine accidental cut details will be on the edge of pixelization problems with most cameras. And you will be hard pressed to find a printer capable of making full life-sized images without also having either noise problems or blocking between adjacent dark areas. If the lens is not a top-quality lens, there will be difficulty due to its MTF drop-off as well.

RESOLUTION CAPABILITY OF A CAMERA

To understand the resolution capability of a camera, it is best to visualize how the pixels in the camera interact with the original object to produce an image. Figure 6.5 shows two bars of a three-bar chart. It is composed of ink printed on paper. The bars and the spaces between the bars are the same width. We will work with a camera with a given lens and we will take photos at various distances. In each case it will be necessary to carefully focus the lens. Also, we will assume that the bars are vertical and align perfectly with the columns of pixels in the sensor chip. A very unlikely circumstance, to be sure, but for this exercise it makes it easier to see the relationships. To get started, set the original simple three-bar chart very close to the camera so that the bars take up one-third of the width of the frame. The test target is three inches wide, so the bars portion is one inch wide. Assume that the camera is now five inches from the test target. For simplicity assume that there are 3000 pixels in each row on the sensor chip. In this situation, in the bars section, all three bars and the intervening spaces will be imaged by 1000 pixels. There are five distances of interest in this case: three dark bar widths and two intervening white spaces. So, each distance will be imaged by 200 pixels. That is, there will be 200 pixels for each bar and each space. The camera will have no trouble seeing the bars as separate and distinct.

Now for the second situation, we will move the test target quite a distance away—in fact, 1000 times further from the camera, or 5000 inches (416 feet, 8 inches). At this distance, the whole section of bars and spaces is covered by a single pixel. Recall that nothing finer than a pixel can be recorded. So no bars are seen by the camera. The frame is 3000 inches wide

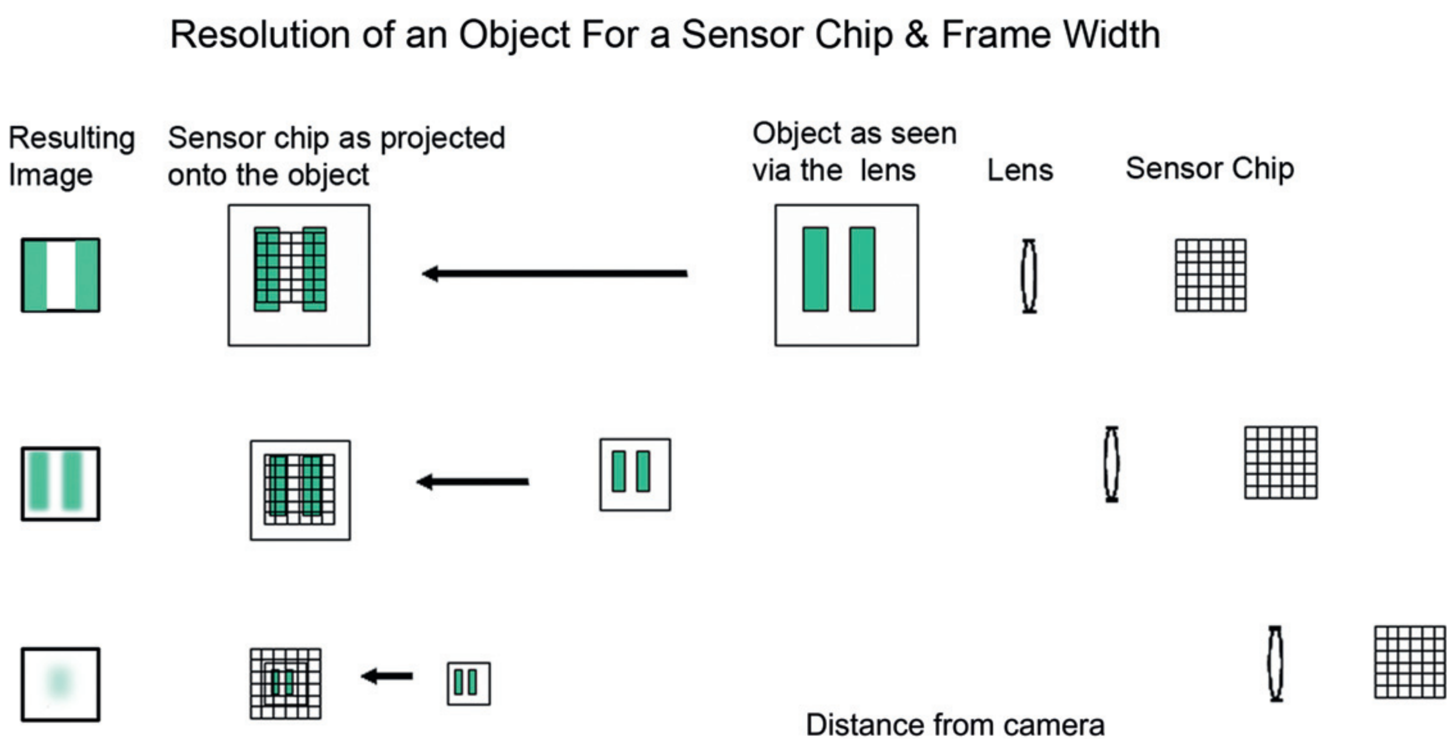


FIGURE 6.5 *Resolution of an Object for a Sensor Chip and Frame Width. The degree to which a digital camera can render detail in an object depends on the number of pixels and the width of the frame at the plane of the object. The sensor chip is shown on the right. To the left of the lens is the object as it would be seen by the chip due to distance of the object from the camera. To the left of this is a representation of the chip as it would project through the lens onto the object at its apparent size. The leftmost column show a rendition of what the resulting image would be like. Up close and the bars are well reproduced.*
At some intermediary, critical distance, the bars are distinguishable but blurry. At greater distance there is just a faint blur.

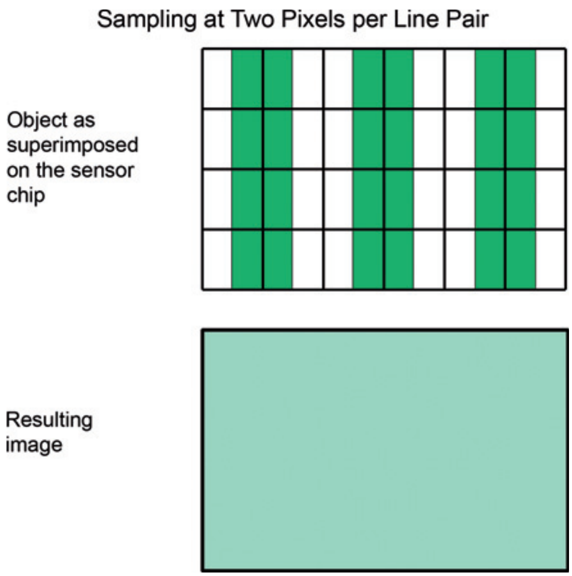


FIGURE 6.6 *Sampling at Two Pixels per Line Pair. The common assumption is that only two pixels are required per line pair. This will work if the lines are well lined up with the columns of pixels. If not, the approach will disintegrate to the point where the lines are aligned with the junctions between the columns of pixels, shown at the top. Since each pixel sees half line and half background, and since the pixels average what light falls on them, each pixel will see the same thing—the average of the overall frame. This is shown at the bottom.*

(250 feet). We have shown that at a distance of five inches from the camera, the image frame is three inches wide and the camera can easily resolve all three bars and both spaces. And, when the target is 5000 inches from the camera, the frame is 3000 inches wide and it is totally impossible for the camera to see any of the bars or spaces. It sees only a gray that is a bit darker than the white background. Somewhere between those extremes there is a point where the camera just resolves the bars and spaces, and beyond that point, the bars are no longer separate and distinct. The frame width associated with just resolving the bars is called the *critical frame width*. The question is, where is that point? If we know that, we will be able to determine what width bars and spaces can be resolved for a given frame width.

There is a branch of mathematics known as Sampling Theory, and one of the key tenets of this field is the Nyquist Sampling Frequency. This holds that if we take two samples per wavelength, and reconstruct using the sinc function, we will achieve adequate sampling. This does not hold for digital photography. The Nyquist requirement would amount to one pixel per dark bar and one per white space. It would also require that the printing system utilize the sinc function to make the print. Neither requirement works in photographic applications. Referring to Figure 6.6, note that when the sampling frequency is twice the wavelength (one bar and one space) the bars are resolved only if the one pixel is over a bar and the next one is over a space, or at least close to this situation. If one pixel is over the junction of a bar and its adjacent space, then the next pixel will be over the junction of the next bar and space. The result will be that both pixels will register some mid-level gray value and there will be virtually no reproduction of the separate bars. As for the sync function, no photographic system uses this function.

Note in Figure 6.5, in which there are three pixels per wavelength, the bar and the space will be represented as separate no matter the alignment, and there is no sinc function requirement for the printer. Practical experience has shown that three pixels per line pair is a robust requirement for a digital camera being able to reproduce a line pair of one bar and one of the adjacent spaces. With this requirement known it is possible to calculate the resolution of a digital camera. We will refer to this factor as the *pixel conversion efficiency*; for a wide range of digital cameras it has been found to be equal to three.

The camera just described has 3000 pixels per frame width. Since three pixels are required for robust resolution conditions, the camera is capable of 3000 pixels divided by 3, which is equal to 1000 line pairs per frame width. The test target assumed has bars and spaces of 0.2 inches each. One line pair will be 1 bar plus 1 space, or $0.2 + 0.2 = 0.4$ inches. We require 3 pixels per 0.4 inches. Since there are 3000 pixels, this amounts to a frame width of $3000 * 0.4 = 1200$ inches. If the width of the frame is W_o (in distance units such as inches or millimeters) and the number of pixels per frame width is P_w, then the size (in the same units as W_o) of a minimally resolved line pair, D will be given by:

$$D = \frac{W_o \times 3}{P_w}$$

Note that line pairs per unit length (LPL) is equal to the reciprocal of D, so

$$LPL = \frac{P_w}{W_o \times 3}$$

This clearly shows that as the number of pixels increases, the number of line pairs per unit length will increase, and as the frame width increases, it will decrease.

The level of detail that can be resolved depends upon the frame width, which in turn is governed by the distance from the camera to the object. With a camera that has 4000 pixels per frame width, we would be able photograph a full hand print that is 5×9 inches and still be able to see ridge detail of 6/1000 of an inch. Since ridge detail is about 10/10,000 of an inch, the shot will work well. If the camera had only 2000 pixels per frame width, the same shot would be limited to detail of 12/1000 of an inch with less quality of reproduction of the detail. To get the detail, the photographer would be required to move the camera closer and settle for a smaller frame width.

Taking the lens formulae from Chapter 3, it is possible to perform some additional interesting calculations. We set up to take a photo with a camera with a given focal length, F, and the object frame width is W_o, and the distance from the camera lens to the object is d_o. Putting these into the lens formulae from Chapter 3, the effective width of the sensor chip, W_i, must be:

$$s_o = s_i * (d_o - F)/F, \quad \text{or} \quad w_o = w_i * (d_o - F)/F$$

$$s_o = \frac{s_i \times (d_o - F)}{F}, \quad \text{or} \quad W_o = \frac{W_i \times (d_o - F)}{F}, \quad \text{or} \quad W_i = \frac{W_o \times F}{(d_o - F)}$$

Finally, if the chip has P_W pixels across its width, then the width of each pixel, W_p, is given by:

$$W_p = \frac{W_i}{P_W} = \frac{W_o \times F}{P_W \times (d_o - F)}$$

Most digital cameras have square pixels, so the height of each pixel can be assumed to be the same as the width. The same process can be used again, working in the height dimension instead of the width. Some cameras have octagonal pixels, and so different methods must be used. Video cameras have rectangular pixels in which the width is 90% of the height. Again, modified techniques must be used.

EVALUATING YOUR CAMERA

First you will need a test target. One easy way to get one is to make an image on the computer with 10 lines pairs per inch. Create a new white, gray-scale canvas with large pixels, like 20 pixels per inch. Then use the pencil tool. Add vertical lines: one column of black and the next column of white. It helps to enlarge the image so the pixels are fairly large on the screen. Once about 20 black lines are drawn, send the image to the printer. You should have black and white lines with 10 line pairs per inch on the paper. Measure to be sure. You have made a quick and easy test target.

It is easier if you mount the test target onto some gray cardboard that is larger than the sheet of paper. Now affix the test target to a wall so that it is vertical. The wall should be some neutral, medium color or the automatic exposure system in the camera will be fooled into making the white paper gray.

Check the manual that came with your camera. It will tell you how many pixels it has. Some give the total (e.g., 7.9 megapixels) and others will tell you how many there are across the frame height and width. If you don't have this, you can compute the values for the rows and columns as follows.

The total number of pixels is Pt, the number per row is Pr, and the number per column is Pc. The basic formula is:

$$Pt = Pc * Pr$$

You also need to know the aspect ratio of the camera. This is the ratio of the frame width to the frame height. Some cameras are 3:4, some are 2:3, and others might be 9:16, or even some other number. If you don't know your camera's ratio, take a picture and print it such that it does not fill the full sheet of print paper. Then measure the height and width. Or simply look at the pixel dimensions in the Image Size window when the image is opened in your computer.

In this case, assume that the camera has a 3:4 aspect ratio. Also assume that the camera has 6 megapixels. (In reality, camera manufacturers round off the indicated number, and in this case, the true value is 5,993,240 pixels). These factors given, it is such that:

$$5,993,240 = Pc * Pr$$

and

$$Pr/Pc = 4/3$$

which is to say that,

$$Pr = 4 * Pc/3$$

Substituting back,

$$5,993,240 = Pc * 4 * Pc/3$$

or

$$5,993,240 = 4 * Pc^2/3$$

Doing the arithmetic,

$$Pc = 2,120$$

and since,

$$Pr = 4 * Pc/3$$

then

$$Pr = 2,827$$

So, for the given camera, the P_w is 2,827. A similar calculation will be needed for the camera you wish to test if the value is not given in the manual. With this number you can make some calculations that will be used to set up the camera test. From the formula given earlier, solve for the equation,

$$D = W_o \star 3 / P_w,$$

for the frame width. That is,

$$W_o = D \star P_w / 3$$

The value for D is known; it is 0.1 inch, the distance across a pair of lines on the test target just made. And P_w is obtained from either the manual or the calculation just demonstrated. In the example, this comes out to:

$$W_o = 0.2 \star 2,827/3 = 94.23 \text{ inches or 7 feet 10 inches}$$

On the wall where the test target is mounted, measure 3 feet 11 inches on either side of the center of the test target, and place some sort of mark, like blue masking tape at that location. For completeness, place marks at 2 feet, 6 inches on either side, and also at 8 feet either side.

Put the camera on a tripod and set it at the height of the test target and have the camera pointing at right angles to the wall. Use center-weighted exposure, some moderate f/stop, and be sure to focus very carefully. It might help to have a large flat object on the wall in addition to the test target in order to focus more easily. Do not use JPEG image output. Set the three distances between the camera and the target such that your pairs of markers are just on the edge of the respective frames. You will be taking a picture of an object with a detail size of 10 line pairs per inch at 3 frame widths. One will have the line pairs well resolved, the next will have them just resolved, and in the third the lines will not be resolved at all. Examine your images on the computer screen to verify that you know the camera's performance.

Since most people would rather check a table than run calculations, it is possible to take the process just described and produce a pair of tables. One table shows the detail size that is just resolvable for a variety of frame widths. The other shows the maximum frame width that can be used for a variety of detail sizes. These are shown in Table 6.1. These are for the camera assumed in the example in the previous paragraphs. Some typical object sizes are shown in Table 6.2.

Please note that it is frame width and not lens focal length that is important in determining the relationship between detail size and frame width. If a wide-angle lens is used, the pixels on the sensor chip will be spread over a wider frame, and if a telephoto lens is used, the frame will be narrower. But the same number of pixels will be spread across the frame. The photographer will adjust the camera-to-object distance to get the same frame width. So the result is that the number of pixels per frame width is the same no matter the focal length of the lens.

Table 6.1 *Frame Width Limits for Given Detail Sizes*

Detail Size, D thousandths of an inch	Maximum Frame Width (inches) 6 megapixels	8 megapixels	12 megapixels
1	0.9	1.0	1.4
3	2.8	3.1	4.1
5	4.7	5.1	6.8
7	6.6	7.2	9.6
10	9.4	10.3	13.7
13	12.3	13.3	17.8
15	14.1	15.4	20.5
17	16.0	17.4	23.3
20	18.9	20.5	27.4
25	23.6	25.7	34.2
30	28.3	30.8	41.1
40	37.7	41.1	54.7
50	47.1	51.3	68.4
75	70.7	77.0	102.6
100	94.3	102.6	136.9
200	188.6	205.3	273.7
500	471.4	513.2	684.3

The table shows the critical frame widths for various detail sizes and camera pixel counts. Example: With a 6 megapixel camera and a detail size of 17 thousandths of an inch, the frame needs to be no bigger than 16 inches. With an 8 megapixel camera one can use a frame width of 17.4 inches, and with a 12 megapixel camera, one can go up to 23.3 inches.

LIMITATIONS

The methods discussed in this chapter are designed to give a reliable resolution. They are based on a special test target and are indicative of results that will be found in practice. But there are some additional factors. If there is nonuniform lighting, the results may be different. If the real object of a photo is a light line on a moderately dark background, it will tend to resolve better than expected. Conversely, a light gray line on a white background will not be seen as well. But the process will give a good indication of the frame size and detail size combinations that are likely to be achievable with a given camera.

Table 6.2 *Approximate Widths of Several Items*

	Item	Thousandths of an inch
1	Human hair	4
2	Fingerprint ridges	21
3	Threads on an electrical outlet screw	31
4	Buttons on a men's shirt	315
5	Minor cut on a person's hand	25
6	Major surgery scar	200
7	Medium ball point pen line	18
8	Cracks in the tread of an athletic shoe	3
9	The dot over an "i" in a newspaper	10

The table shows rough estimates of the widths of several items. Used in conjunction with Table 6.1, we can begin to see the number of pixels needed for certain applications. Only the full frame width information is needed beyond this.

UNDERSAMPLING

The array of image-sensitive tiles on the surface of the sensor chip is what captures the light coming from the object. Each tile, or pixel, takes in the light from a small portion of the total scene. It integrates all photons coming to it into a single reading. In combination, they capture the overall scene. If each pixel were in direct contact with its neighboring pixels, there would be no information lost. This is referred to as *critical sampling*. But just as a tile surface has grout lines separating the tiles, sensor chips have spaces separating the pixels, and proportionately, those spaces are wider than the grout lines we see on bathroom walls. The result is that some of the information coming from the image is lost. This is referred to as *undersampling*.

In fact, in most digital cameras, each pixel has a colored filter on its front surface, which means that it sees only one of the three primary colors: red, green, or blue. Then in order to make a complete picture, the missing colors must be estimated from the information that is collected. The estimation process is called *demosaicing*.

The combination of the spaces between the pixels and the demosaicing contributes to the limitation of the camera's resolution. A few cameras and most scanners use three separate chips, so the demosaicing is not an issue. Since the amount of undersampling is well known by the designers of the camera, and the process is well defined, it is possible to use image enhancement filtering to try to reconstruct what the original should be. This is done, but it is at best an estimate. Perfection is not possible.

PRACTICAL OPTIONS

In traditional photography, if the camera and film combination will not give sufficient resolution for a given overall frame size, you can always seek out a higher-resolution film. With digital photography, the chip is an integral part of the camera, and cannot be swapped out. The result is that if the given camera will not render the fine detail needed at the required frame size, either a new, higher-resolution camera is required, or two images should be captured with the camera displaced under half the scene width. Then stitching software can be used to merge the two images into a single image covering a larger area. This process requires that the subject of the image is not moving and that the lens does not suffer from significant barrel or pincushion distortion. It is also important that the camera be the same distance from the object in both shots, and that the lens not be refocused. The optical axis must be perpendicular to the plane of the object. For these reasons, you should be careful in choosing the camera for the application when purchasing. It is possible to stitch together many images, but the optical continuity issues become magnified when this is done.

NOT ALL DETAILS ARE LINE PAIRS

So far in this chapter much of the attention has been given to resolving line pairs. But not all details are line pairs. For example a light-colored thread on a black boot is essentially a single line. So even if the calculations indicate that one line pair per hundredth of an inch is the resolution limit for a given set up, you might easily find that the thread on the boot might show up even though the thread is only one-half of a hundredth of an inch. It may not be clear, but it would be noticeable that something thread-like is there. In other words it is not reliably resolved, but something is apparent.

PRINTER CONSIDERATIONS

More detail on printer technology will be covered in later chapters, but suffice it to say for now that the printer can be the limiting factor in making images with adequate quality to sustain forensic investigations. Printers generally work by placing equally sized bits of ink (dots) or dye on a white paper base. The number of dots per inch combined with the number of dots needed to achieve a pixel will limit the resolution of the printer. For normal viewing of prints, 200 equivalent pixels per inch will generally be satisfactory. High-quality images might be recorded at 300. But if you are looking through a magnifying glass, for example, at fingerprints at life size, at least 500 pixels per inch is recommended. When choosing a printer, a high level of attention must be paid to the needs of the intended application. Finally, if the images will be evaluated on a computer screen, then the printer resolution is not an issue.

EXERCISE

Setting the Resolution in Prints and Cropping in the Camera

This exercise is intended to give practice in measuring the resolution of a digital camera. As with any such exercise, have a pad and pencil at hand and record all settings and data values as you go so that the results can be clearly understood at the end.

To measure the resolution of your camera, start by finding out how many pixels your camera has. Some cameras list only the total pixel count in megapixels, whereas others, in the users manual, give the number of pixels in each row and column. Not all the pixels are actually used to take pictures, however, so you might want to test for yourself. The easy way to test is to take a picture, open it in an image editing package such as Adobe PhotoShop software, and look at the pixel dimensions of the image—it will give you the row, column, and total value directly.

If the manual gives usable pixels per row and column, you are all set. Just record the number of pixels per row assuming square pixels. If the pixels are octagonal, you will have to experiment directly.

If it gives only total megapixels, you need to know the aspect ratio—the width of the frame divided by the height. Most cameras have a 4:3 aspect ratio. Now you can convert the total pixel value to rows and columns as follows:

1 Use the following equation to estimate the number of pixels in a row:

$$Pw = \sqrt{(4 * M)/3}$$

where Pw is the pixels in each row, and M is the total number of pixels (it is normally expressed in megapixels, so be sure to multiply that by one million before making the calculation).

2 Use the following equation to estimate the number of pixels in a column:

$$Pc = (3/4) * Pw$$

where Pc is the number of pixels in each column and Pw is the number in each row, as just calculated.

3 Multiply the row and column values and check against the total to be sure you have it right.

As a test target, use the vertical bars target on the web site. Download that image and print it. The fine bars are 20 per inch, in the next size up the bars are 10 per inch, and the largest ones are 5 per inch.

It is best to concentrate on the 10 bars per inch section of the target. If the bars are 10 per inch, then the size of the detail you are seeking to reproduce, D_o (detail on the object), is 1/10 of an inch. Given this value and the value for P_w, you can calculate the critical frame width, W_o (width of the frame at the focal plane of the object).

$$W_o = (D_o * P_w) / 3$$

You are now ready to make your measurements. Take the test target and tape it to a wall about four to five feet off the ground. Make an X out of two 3-inch strips of blue masking tape on the target, but not covering any of the bars. You will use the X later to focus the camera. Try to make the bars as vertical as possible. Measure a distance equal to ½ W_o on each side of the center of the set of 10 bars per inch target and place a piece of blue masking tape at both ends. The distance from one piece of tape to the other should be W_o inches.

Place the camera on a tripod. Set the tripod close to the wall and adjust the height of the camera such that the camera lens is at the same height as the 10 bars per inch target. Set the camera to spot metering and aperture priority. Dial in an aperture of about f/5.6 or f/8 or something like this. Set the camera to output to RAW files or JPEG images at the largest and highest-quality setting. Now, following a line perpendicular to the wall, move the camera back from the wall until the bits of tape are at the extreme ends of the frame of a picture taken by the camera. Set the camera to manual focus and be sure that the camera is in focus by using the X you placed on the chart. Try to make the camera base parallel to the floor (assuming a level floor). The idea is to make the vertical bars line up with the columns of pixels as much as is reasonable. Pictures taken at this setting should just resolve the 1/10-inch bars. You might want to experiment by taking a few shots six inches closer and six inches further away to bracket the distance setting. You will want to use enough light to have a fairly short exposure time and a mid-scale f/stop.

Measure and note the distance from the wall to the camera. Move the camera toward the wall such that it is now half the original distance from the wall. Refocus and take some additional pictures. Again, bracket by about six inches on either side. The frame width here should be half the critical width for 1/10-inch bars and close to the critical width for the 20 bars per inch target. Finally, move the camera away from the wall so that it is now twice as far from the wall as the original setting. Refocus and take some more pictures. Bracket this setting as well, but move one foot either way at this distance. This setting should be close to the distance for just being able to resolve the five bars per inch lines.

You should now have photos of the test target taken at the critical frame widths for 5, 10, and 20 line pairs per inch. In addition there are some a bit wider and more narrow than the calculated values. You will want to examine all of these on your computer. Open the images, ignoring any with clear problems, and select images at each of the nine settings for examination. You should have three that are close to each of the three critical settings and you should examine the image to see if the calculations held true. Since the test target is black and white, it may be easier to convert the images to grayscale for viewing. Otherwise you may see some color aliasing, which might be distracting.

So for the images with the narrowest width (closest distance) you should see all the lines of all the three line widths such that you can clearly tell that there were separate lines in the original. The line edges might be blurry, but the lines are distinct. If there is a diagonal pattern, that might be because the pixel columns and the bars are misaligned. It would be nearly impossible to get them perfectly aliened, so take the best set. See if the images from the bracketing make any noticeable difference.

Repeat the examination of the pictures from the next distance, the one critical for 10 lines per inch. You should see the lines from the charts for five and 10 line pairs per inch, but not the 20. Finally, repeat the process for the widest frame width. Here you should see only the five line pairs per inch part of the target. All the others should merge into gray patches instead of showing distinct lines.

Chapter 7

Color Space

Some simple truths regarding the world around us: "Roses are red, violets are blue…", the sky is blue, and leaves are green. All are true but not sufficiently accurate to take good photographs, or print a magazine, or decorate a room. A bit more precision and consistency is required. Some other common assertions: a person's foot is a foot long, from the nose to the end of the outstretched fingers on an extended arm is a yard. But in order to design machines, more precision and consistency was required. So the ruler was invented. But measuring distance with a ruler is much easier than finding a suitable "ruler" to measure colors. A distance is a physical thing and the effect on the observer is not really that significant in most instances. Not so with color.

Color is a perception. It is the conclusion your mind comes to when stimulated by a certain stimulus under specific conditions. Change the conditions, and the same stimulus will appear to change color. And there are many different stimuli that can be perceived as having the same color. The result is that complex measurements and systems are required.

There are five main elements to consider:

1 **R**—the response. This is a value, and in the case of the perception of a color, it is the impression the person has of the color. If it is a light meter, it is the reading on the meter. If it is the sensor chip in a digital camera, it is the value that that portion of the chip returns.

2 **L**—the existing light. This is in the form of the intensity (photons per area per second) of that light for each wavelength in the visual spectrum. Under daylight, dark blue and black are easily distinguishable from each other. Under sodium arc light, they are very hard to distinguish. You can easily wear navy socks and black pants to a fancy party if the light is from incandescent lamps.

3 **O**—the object in question. We want to know the color of a particular thing so that when we represent it in a photograph, it will look like it really looks when seen directly. The socks are really blue and the pants

are really black unless you see them under certain lighting conditions. This will be measured as the proportion of incoming photons per square centimeter per second that arrive on the surface and are reflected off of it. If it is a transmission object, such as a film negative, it is the proportion of the photons that actually penetrate and come out the other side.

4 **F**—any filters in the system. When you sit in church in the morning as the sun comes up and shines through the stained-glass window onto the white walls inside, the various areas of the wall take on the colors of the bits of glass. The wall is still "white," but if you did not know that from other visits to the church, you might think there was a painting there. This is measured as the proportion of the photons per square centimeter per second that actually penetrate the filter and come out the other side.

5 **S**—the sensitivity of the sensing element. A person's eyes have no sensitivity to infrared light photons, but silicon sensor chips have a lot of sensitivity to them. If you walk into a room that is bathed only in infrared light, the room will be dark and you will not see anything. But your video camera with infrared sensitivity will have no trouble taking ghost-like movies. The sensitivity is measured as the output per incident photon. There is an assumption that the incident beam of photons is allowed to strike the surface for a specified amount of time. The output will depend on the sensor and where we choose to measure. For example, it could be the charge accumulated in Coulombs per incident photon. Or, when applied to a capacitor, the voltage achieved per incident photon. Or, when converted to a number by an analog-to-digital converter, the number obtained per incident photon. In any event, this is dimensionally the same as the value of R from factor 1.

The basic idea is that factors 2 through 5 are combined to predict factor 1. This was introduced in Chapter 4. If factor 1 is expressed as a color, then it will be necessary to repeat the process at least three times and then combine the three readings.

The process for combining the four independent factors requires that their values be determined for very narrow bands of wavelengths all across the relevant spectrum. As the wavelength bands become narrower, the number of these bands will go up. Theoretically, if the width approaches zero, then the number approaches infinity. The next step is to take the first band, ascertain the values for all four factors at that wavelength, and multiply the factors together to get the combined effect at that part of the spectrum. Then the next band is evaluated the same way and the results for the two bands are added together. The process is repeated for each band all the way across the spectrum and the result is the predicted response, R. In other words, we will integrate the products of the factors by wavelength across the relevant spectrum:

$$R = \int_{\lambda} (a \times Light) \times (b \times Object) \times (c \times Filter) \times (d \times Sensor) d\lambda$$

Note that a, b, c, and d are constants that are employed to ensure that there is consistency across units of measure. Note that if the light source in a particular set up is intense in the blue (shorter wavelengths) portion of the spectrum, but the sensitivity of the sensor is low in the blue, the two will offset each other, and the response will be low. Conversely, if the light were rich in red (longer wavelengths) and the sensor were also very responsive there, the two would reinforce each other to give a high response.

Human Color Vision

The human visual system starts with the eyes and extends back into the brain. In fact the optic nerves, which connect the eyes and the occipital lobe in the back of the brain, appear to really be a part of the brain itself. Indeed some argue that the eyes are merely an extension of the optic nerves and the optic nerves are an extension of the brain itself. A photographic system is supposed to provide a fair and accurate representation of what we would have seen had we been at the original scene. That is to say, the system must be such that it can capture the information at the scene and then represent it in such a way as to render a recognizable picture to a human viewer. This discussion will start with a description of certain key elements of the eye and briefly discuss the processing of the image by the brain.

THE EYES HAVE IT

Each eye is a ball on the end of a hose. The ball is the famous eyeball, and the hose is the optic nerve. The optic nerves—one from each eye—cross mid-skull and attach to the occipital lobes on the back of the brain on the opposite side from the attached eye.

The eye is a complex and magnificent organ, shown schematically in Figure 7.1. It is about 1 inch, or 2.5 centimeters, in diameter. The optic nerve exits from the rear, and the opening that allows light to enter the eyeball is almost opposite this point. Covering the eyeball opening is a protective window called the cornea. This covers the lens. The lens is attached to a set of muscles that are used to adjust the focal length of the lens. The muscles allow the lens to be fat and rounded (short focal length) or more flat (longer focal length), and they can also help adjust the placement of the lens relative to the back of the eyeball, thereby adjusting the focus. Behind this is a muscular assemblage referred to as the iris. The iris can open fairly wide

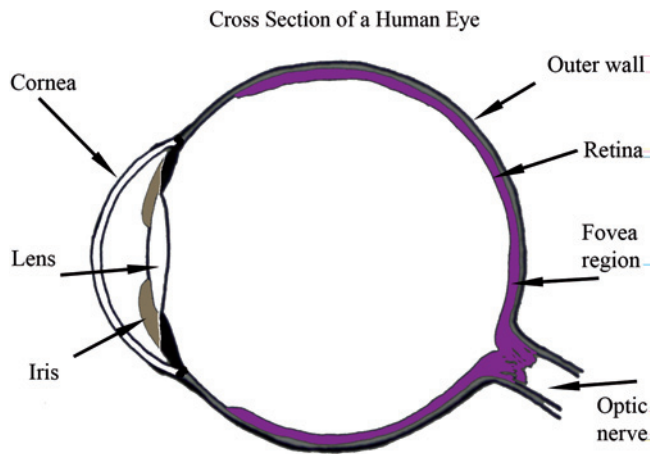


FIGURE 7.1 *Schematic of the Human Eye. Shown is a cross-section of a human eye, indicating the locations of the key elements of that organ.*

to allow more light to enter, or shut down somewhat to allow only a small portion of the incident light to enter. In the center of the ball is the vitreous humor—a clear, viscose fluid that fills what otherwise would be a void.

Coating the back of the eye is the retina. This marvelous system is comprised of millions of sensor cells. Some are shaped like cylinders and are called rods, and others are shaped like cones and, appropriately, are called cones. Rods and cones are the light-sensitive elements in the eye. They absorb photons and emit nerve impulses accordingly. Higher numbers of photons per second result in stronger responses. The eye can respond to as few as five to 10 photons and as many as a few million.

The spectral sensitivities of the elements of the human eye are shown in Figure 7.2. Rods have a higher sensitivity to light than cones—one that peaks at about 555 nanometers (nm)—but has sensitivity well into the blue and the red. The rods are the primary detectors in low light. The response of the rods in low light is called the scotopic luminosity function. The cones have lower sensitivity but come in three varieties, often called the red, green, and blue cones. Some cones are sensitive in the red portion of the spectrum, some in the green, and some in the blue. The response of the eye in brighter conditions is comprised of components from both the rods and all three cones, and is called the photopic luminosity function.

Between the vitreous humor and the sensor cells in the retina there is a network of virtually transparent, interconnecting nerve cells. The rods and cones actually face the back of the eyeball. These interconnecting nerves combine the impulses that come from the individual rods and cones and start to process the raw data into an image. The combined information is then routed to the optic nerve. The optic nerve adds a bit more processing of the impulse information and transmits the partially processed data to the visual area of the occipital lobes of the cerebral cortex. It is here that

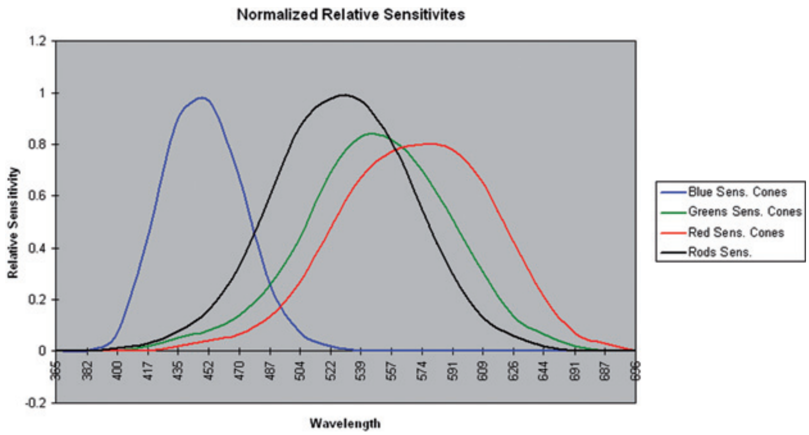


FIGURE 7.2 *Human Eye Sensitivity. The sensitivities of the red, green, and blue sensitive cones in the human eye are shown normalized to the areas under the curves being equal to one. The sensitivity of the rods is shown with its peak sensitivity set to one.*

the image that you "see" is formed. The lens in your eye creates a projection of the scene before you on the retina, but the image that you actually "see" is the virtual version of this rendered within your brain in the back of your head. When you look at someone who is looking at something, you can see that his eyes actually dart around a bit. Also he will blink from time to time—not a conscious blink, but an automatic one. Despite the movements of his eyes, he will see a steady image of what is before him. In fact the brain is combining the various inputs from the eye to create the cerebral image, and it is the cerebral image that he sees and of which he is conscious.

The muscles that hold the eyeball in place also allow it to move. The muscles in your iris allow you to adjust the amount of light coming into the retina. The muscles in the lens allow you to focus at various distances. And all these are very active while you are looking at something. As your eye moves, adjustments are made for local brightness variations, and for distance. The result is not a single image, but several separate images. Your brain compiles these into the single image that you see. There are many ways in which all of this is manifest, but you are so used to how automatically this is all done, that you are unaware of the process. For example, the spot where the optic nerve exits the eyeball is actually blind. There are neither rods nor cones there. And it is nearly in the center of your field of view. You are totally unaware of this blind spot. Unless you stare at something long and hard, you will tend to have everything in the field of view in focus. This is because when you looked at each particular spot, your eye focused to the appropriate distance. So the composite in your brain has everything in focus. You can also see details in very bright portions of the scene as well as in very dark areas. This is because when you looked at these spots, your iris made the required correction. Next time you are tested with an eye chart, note the finest lines you can read with the eyes one at a time and then in combination. You can see considerably better with both eyes than you can with them one at a time. Just ask your eye doctor. And, all this happens in the blink of an eye.

The retina has cones tightly packed into a small area near the center of the field of view, the *fovea* or *macular*. This area of the retina allows for good color vision. In areas further from the fovea, the predominance of cones drops and you see primarily with rods. Also the density of sensor cells is somewhat lower altogether. The result is that peripheral vision is less sharp and primarily comprises outlines. If you need to know about something in a peripheral area, your brain will move your eyes or even turn your head to get the portion of the scene into the fovea.

The rods and cones actually can move relative to each other. So if it is dark, the rods move forward, and if it is bright, the cones take the front. When you go from a bright setting into a dark one, this transition must occur and it may take a few seconds. The phenomenon is called *dark adaptation*. The reverse occurs when you go from a dark area to a bright one—*light adaptation*. The dark-adapted eye is extremely sensitive, but unless elements of the scene are bright, they are seen as shades of gray—rods do not distinguish among colors.

The movies projected in a theater are really composed of a series of still images shown in rapid succession—24 frames per second. Televisions show an even more complex series of stills. Standard broadcast TV breaks movies into frames and then breaks the frames into fields. It also breaks the frames into a series of stripes across the image, each becoming a row of information. The odd rows comprise one field and the even rows comprise the other. The TV uses 1/60th of a second to show the odd field, and then in the next 1/60th, it shows the even field. This is followed by the next frame. And so on. You see this as a movie with continuous action. The phenomenon that allows this to work is *persistence of vision*.

The rods and cones in your eyes hold a chemical called *rhodopsin*. The photons are absorbed by this chemical and the result is that an electron is freed and a nerve pulse is formed. The rhodopsin is good for only so many reactions and then it has to be replaced. Your eye does this handily, making as much chemical as required. But it takes about 1/16th of a second for the flushing and replenishment. This means that if the scene changes slightly and more rapidly than this, you will not see the transition. You will see one state and then the next. If you look at a red light for several seconds, and then quickly look at a gray screen, you will see a cyan after-image in the shape of the red light. This is because you used up all the red-sensitive rhodopsin and until it is replenished, you will not be able to see red. Subtract red from gray and you have cyan.

We can look at a representation of the spectrum of colors and convince ourselves that we can see all these colors. Yet we have sensors for only red, green, and blue light. And the red and green sensitivities overlap significantly. So how can we see all the colors? For example, we can see yellow, but we have no yellow sensor. We can also see brown, but there is neither a brown sensor nor a brown wavelength. So how do we see all the colors? The answer is that our eyes pick up sensations in groups of three. Responses are picked up by our red, green, and blue cones simultaneously and the inputs are combined by our brains to determine the color of the object we are viewing.

For example, if we're looking at a computer screen and there is a portion that is sending out roughly equal amounts of nearly superimposed red and green light, we would see the patch as yellow. If we had made a print of that particular patch and compared it to the screen, we might see the two patches as very similar in color. But there is a catch. The computer screen has the ability to produce red, green, and blue light—and only those colors. It makes us think we are seeing all the other colors simply by superimposed presentation of selected amounts of those colors. So to make the yellow become orange, we would increase the relative amount of red in the mixture, and to move toward chartreuse, we would increase the relative amount of green. To get brown, add in a bit of blue. And so on. What you see is in your brain and what exists in the external world is a different thing entirely.

As for the print, the story is quite the reverse. Let's assume that the printer was a dye sublimation printer. These devices can produce only cyan, magenta, and yellow. They fill in all the other colors by presenting varying amounts of these colors. To make the patch more orange, add more magenta, and to make it chartreuse, add more cyan. To make it turn brown, take out some of the yellow.

Summing it up, in the one case, we are looking at a combination of red and green light and seeing yellow; and in the other, we are looking at yellow and seeing yellow! To take it a step further, if the yellow patch were surrounded by white, the yellow would seem a bit pale, but if it were surrounded by blue, it would seem more vibrant. Remember that we see in our brains. Color is a psychological phenomenon, not just a physical one. The spectrum that we often see portrayed with associated wavelengths is a bit of a myth. The phenomenon of wavelength is physical and physically related to the energy of the photons by the laws of physics. The fact that when people look at light with a wavelength of 540 nm and call it green is a psychological phenomenon. Color science must utilize the physical to predict the psychological. Photographic science takes the desired psychological result and works back through the physical to produce that desired result, given the constraints of the outside world.

When testifying and presenting images, there is the inevitable question: "Is this a fair and accurate representation of the scene?" This does not mean that it is a physically accurate representation, but rather it is what an observer at the scene would have seen and it does not distort the facts of the scene in ways that will lead us to draw an inaccurate conclusion regarding what was at the scene. So, if a yellow patch is seen on a computer screen and it shows up as yellow in the image, that is a fair and accurate representation, even though the computer screen picture was actually putting out red and green in roughly equal amounts (it cannot put out yellow *per se*).

COLOR SPACE

Assume that we have found a shell casing on the floor of a rectangular room, and want to describe exactly where the item was found. This could be done easily by measuring the distance from one of the shorter walls and parallel to the longer wall and calling it "L"; and the distance from the longer wall and parallel to the shorter wall and calling it "W." With the starting walls duly identified, we can locate the casing by the ordered pair of numbers, L and W. The floor becomes a two-dimensional space in which the dimensions are length and width. If we had a floor lamp there and wanted to locate where the bulb was, we would add the height of the lamp, H, and now we have a three-dimensional space with dimensions of length, width, and height. In this three-dimensional space we can locate any point by means of an ordered trio of numbers, usually separated by commas: l, w, h.

Now let's consider a different three-dimensional space: music. Each note has a primary pitch, loudness, and a number of overtones (all tones except that at the primary pitch). If a note has a high pitch, several overtones, and is loud, we would call it brassy. If it were the same pitch and loudness, but only a few overtones, we would call it shrill. Low pitch tones with few overtones and low loudness are somber, and if the loudness is increased, it becomes bombastic, and so on. Note that knowing the pitch says nothing about the number of overtones or the loudness. Likewise knowing the loudness says nothing about the pitch or the overtones. And knowing the overtones says nothing about the pitch or loudness. The important feature is that the three dimensions are independent of each other, or *orthogonal*. Also, each sound that we hear can be put in a particular position in the three-dimensional space and any note from a given location will have known pitch, loudness, and number of overtones. In order to engineer photographic systems, we need a space that allows us to take any small element of a picture and assign it a location and know that any color assigned to that location is the same as any other in that spot. We don't want any dimension to imply any of the others, and it would be very nice if we could attach a name to each location. Such a coding system is known as a *color space*. This allows us to represent a color as means of a set of numbers—this is critical to any process in digital imaging.

In order to have a workable mathematical space there must be a clear definition for each of the dimensions and there must a clear set of rules for how the values will be determined. In some cases the dimensions are defined by the primary colors, and in other cases by theoretical constructs. The rules must be based on mathematics. The most common color space arrangements are either in three-dimensional Cartesian coordinates, or in two dimensions, which have one axis that is Cartesian and the other two that are in polar coordinates.

THE RGB FAMILY OF COLOR SPACES

Since the eye has three color receptors, we know that we will need at least three dimensions to the space. The simplest color space has the dimensions of red, green, and blue—just as the eye does. We will need to define the meanings for the terms red, green, and blue. This is done by establishing a particular filter and light source. This, in effect, gives the spectral distribution of each primary color. It tells us (at least in relative terms) how many photons there are per square meter per second at each wavelength. Once this is done, we can take any color sample, measure how much red, green, and blue light is coming from it, and assign it a location in our space. This space is called the RGB color space. There are a number of working RGB color space designations within the RGB umbrella and they differ in terms of the spectral components of the specific primaries. To determine the space, we use the equation for system response cited earlier and select a particular light source, filter, and sensor response. With those three controllable factors determined, each image sample has a unique response. If we use the red filter, it tells us the R component. Substituting the green and blue filters will yield the G and B components, respectively. The values for R,G, and B are the dimensions of the space, and the combination of light source and filters employed defines the Primaries, or primary colors.

We can now assign that element a location in our color space and we know that any element coming from that location has that particular color. We have just defined the RGB, or Red, Green, Blue color space. Note that if different filters are used, we have a variant of the same space. Modern digital technology has spawned a whole family of RGB working spaces. They use different filters and are designed for different specific systems. For example: Adobe RGB, the Kodak Pro Photo RGB, and sRGB are in widespread use today. These can easily be converted from one to another by use of a set of three linear equations.

If large amounts of red, green, and blue light are added together in roughly equal proportions, the result is white light. The absence of all three is black. Figure 7.3 contains a cube structure showing schematically the locations of the colors in the RGB space. Note that black and white are at opposite corners of the cube and all the other colors have a particular location inside the cube. In typical digital systems, each dimension has 256 values, which is 2^8, and works well in the binary number system. In this case there are $256 \times 256 \times 256$, or 16,777,216 different colors. This is referred to as 8-bit color or 24-bit color. Newer equipment can work at greater bit depths, for example 12-bit color or 16-bit color, resulting in vastly larger numbers of colors.

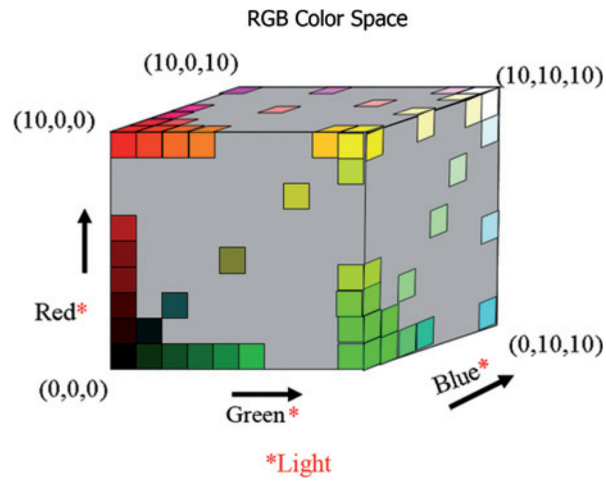


FIGURE 7.3 *The RGB Color Space. This shows a simplified diagram of how the RGB color space is laid out in three-dimensional Cartesian coordinates. Note that the blue axis is really on the bottom-left corner of the cube, and not visible in this drawing. The bottom-front, left corner is the origin, which is black since there is a zero amount of each light. The upper-rear, right corner is white, full amounts of each light.*

There is reason to believe that the eye can see up to 16 million colors, and so adding more colors is more for arithmetic reasons than visual ones.

The RGB space (or some particular form of it, to be more precise) is used for systems in which we are adding light. If there are three lasers—one red, one green, and one blue—all set to illuminate a portion on a screen at the same time, we are adding light. A computer screen has very small lighting elements. Each one can glow either red or green or blue, and the intensity for each color is controlled by the device. The elements are so small that the eye cannot see them as individual elements and so the light combines into a single beam for all intents and purposes. Thus the colors of light are added together. The important feature is that when using a particular RGB space, we are adding together amounts of light of the three colors. So the RGB space is *additive* and the primaries are the *additive primaries*.

THE CMY COLOR SPACE

If we add red and blue light together in roughly equal proportions, the resulting mixture is a purplish color called magenta. Adding together red and green results in yellow, and adding together blue and green results in cyan, an aquamarine-like color. Another way to view this same information is to notice that magenta is characterized by the absence of green. It was made by adding together red and blue but no green. So magenta is the negative of green. Likewise yellow is negative blue and cyan is negative red. In the RGB space, we add selected amounts of light of the colors to black to get the desired color. In the cyan, magenta, yellow—or CMY—space, we start with

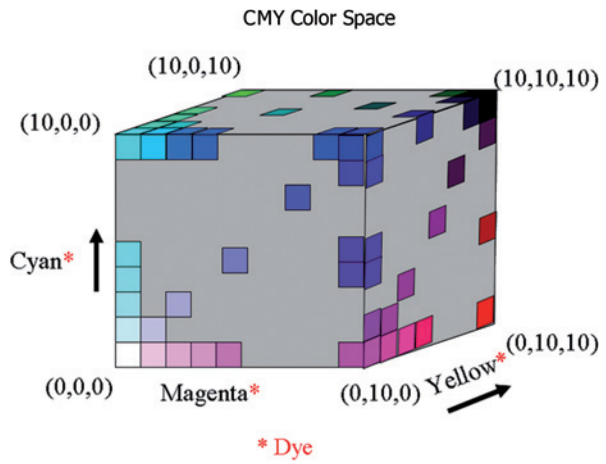


FIGURE 7.4 *The CMY Color Space. This shows a simplified diagram of how the RGB color space is laid out in three-dimensional Cartesian coordinates. Note that the yellow axis is really on the bottom-left corner of the cube, and not visible in this drawing. The bottom-front, left corner is the origin, which is white since there is a zero amount of each dye and the full white light comes through. The upper-rear, right corner is black, full amounts of each dye.*

white light and use different-strength filters colored cyan, magenta, or yellow to subtract different amounts of the opposite colors. Hence, the CMY space is known as *subtractive* and the colors are the *subtractive primaries*.

The CMY space is shown as a cube in Figure 7.4. Note that the origin is white and the opposite corner is black—just the reverse of the RGB space. Note that again the fineness of the scales is adjustable. In 8-bit color, there will be 256 steps, in 12-bit there will be 4096, and in 16-bit there will be 65,536. Respectively, the number of colors in each of these spaces will be 16,777,216 colors, 68,719,476,736 colors, and 281,474,976,710,656 colors.

The CMY space is used for situations where the light source is not a part of the display. For example, a paper print is viewed under ambient light. A transparency is viewed in a projector by means of the projector's lamp. The point is that we start with white light and then subtract selected amounts of light of within certain wavelength bands by means of varying the amounts of the three dyes (or filters): cyan, magenta, and yellow. The result is that the viewer is allowed to see the red, green, and blue light that remains.

A variant of CMY is CMYK, where the K stands for black. Note that equal amounts of cyan, magenta, and yellow dye result in gray. This is equivalent to the same amount of neutral or black dye. That is, 0.3 density units of cyan in combination with 0.3 density units of magenta dye, in combination with 0.3 density units of yellow dye is equivalent to 0.3 units of black dye (note the equivalence is not exact, but easily predicted). A patch of color comprising:

0.3 units of cyan dye, 0.5 units of magenta dye,
and 0.6 units of yellow dye

is equivalent to:

0.0 units of cyan dye, 0.2 units of magenta dye,
0.3 units of yellow dye and 0.3 units of black dye

The CMY system would use a total of 1.4 units of dye (0.3 + 0.5 + 0.6). The equivalent CMYK patch has a total of 0.8 units of dye (0.0 + 0.2 + 0.3 + 0.3). Remember from Beer's Law the optical density is proportional to the amount of actual dye in the medium. The result is a savings of 0.6 units of dye. In large printing jobs, where the cost of ink is significant, or in making multiple prints on a color printer, the difference can add up quickly. As a result, the CMYK system is used widely in printing applications.

It is important to note that the K dimension is not really a fourth dimension. Instead it is derived from the C, M, and Y channels and so the CMYK system is really a three-dimensional system. And because of the simple relationship among the primaries in the RGB and CMY color spaces, it is easy to convert from one to the other.

In summary, the additive and subtractive primary color spaces are intuitive and work with primaries that are recognizable colors. They work well for identifying colors and for making basic calculations. The problem with them is that they do not react the way that people actually see. For example, if you had an RGB color patch but thought it was too light, you could reduce the amounts of red, green, and blue light in that patch in the hope of making it darker. The problem is that it requires three separate changes to do that and the resulting patch will probably appear to have shifted in hue (basic color) and possibly changed in its vibrancy at the same time. Through repeated adjustments you will eventually get the result you sought, but it will not be easy. And if you changed the whole image instead of only one small part of it, the rest of the image might have changed in curious ways.

THE HSL AND HSV COLOR SPACES

To find a color space that behaves more like how people see, the Hue, Saturation, Lightness (HSL) and Hue, Saturation, Value (HSV) color spaces were developed. Both of these have, as two of the three required dimensions, hue and saturation. Hue refers to coloration. Imagine all the colors arrayed in a circle. In the center is white—a combination of all the colors—and at the periphery, the colors are highly pure—that is, they are composed of one or two of the additive primaries, but contain none of the third. Hue refers to the angle at which a given color is located. Saturation is given by how far the color is from the center. White is fully unsaturated, and the colors at the periphery are fully saturated. The two systems differ in the third dimension.

In the HSV space, shown in Figure 7.5, value is an indicator of brightness ranging from zero to one. At zero there is no light and no color—it

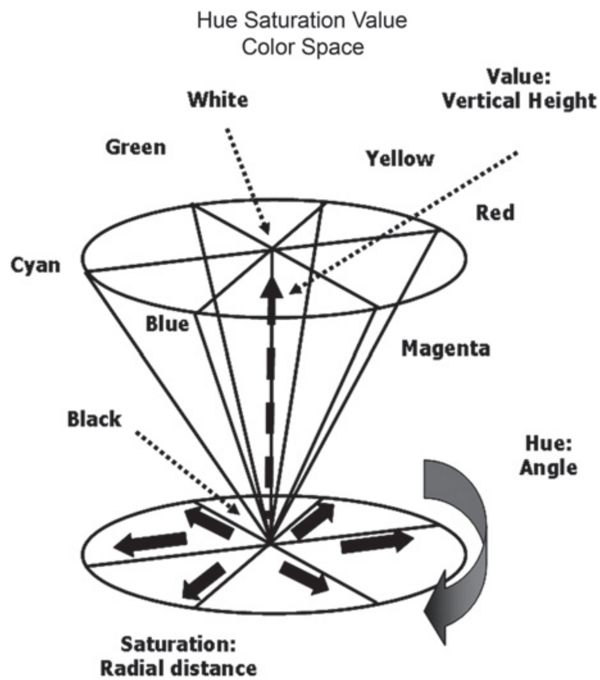


FIGURE 7.5 *The Hue Saturation Value Color Space. This shows that the black-to-white axis runs vertically up the center of the diagram and the colors lay in rings around this axis and can be located in polar coordinates. Note that zero saturation colors lay on the vertical axis.*

is black. Saturation and hue have no meaning. So at zero value, we have a point: the point of a cone. As value is increased, the cone grows in diameter and colors appear, and they can increase in saturation as they move to the edge of the circle. At value equals one, the center of the circle is white and the colors at the circumference are very bright and very pure. Overall, this space is shaped like a cone. The value dimension runs from the black point at the bottom to the white spot in the center of the circle at the top. The colors are arrayed around the circles at each value, with the center being desaturated (running from black, through grays, to white) and very saturated at the edges of the circles.

In the HSV space, the three dimensions can be directly calculated from RGB values and vice versa using simple linear equations. If we were to darken an image by decreasing the value, the only other change that would appear would be that the saturation limits would decrease. In other words, as the scene gets darker, colors are less vibrant (as well as darker). After the adjustments were made, sending the image to the printer would not be a problem as the CMY or CMYK values would all be computed according to the one change that was made. If the colors in the image were too dull, we could increase the saturation and that is the only change that would occur. Again, the CMY values would all be adjusted accordingly. If the colors were off a bit, changing the hue would shift not only the key colors being examined, but all the other colors at the same time. The value and saturation would not change.

However, HSV has one issue of contention. In real life, as patches get very bright, they become white and lose any individual hue. This is where the HSL color space comes in. This is like two cones back to back, as shown in Figure 7.6. The bottom cone is very similar to the HSV cone; the top cone is inverted but is white at its tip (on the very top) instead of black.

In the HSV space, hue is still noted by an angle around the circumference of the circle, saturation is still denoted by distance from the center of a given circle, and value replaces lightness and runs up the center of the figure from the black tip at the bottom to the white tip at the top. The most saturated colors are found where value is equal to 0.5—the vertical middle of the figure. As color patches get brighter than 0.5 in value, they lose saturation. Likewise as they get darker than 0.5 in value, they also lose saturation. The hues are consistent all the way up and down the radial arms of the figure.

In summary, the HSV and HSL color spaces behave more like the way we see. They have primaries that are neither intuitive nor are they colors.

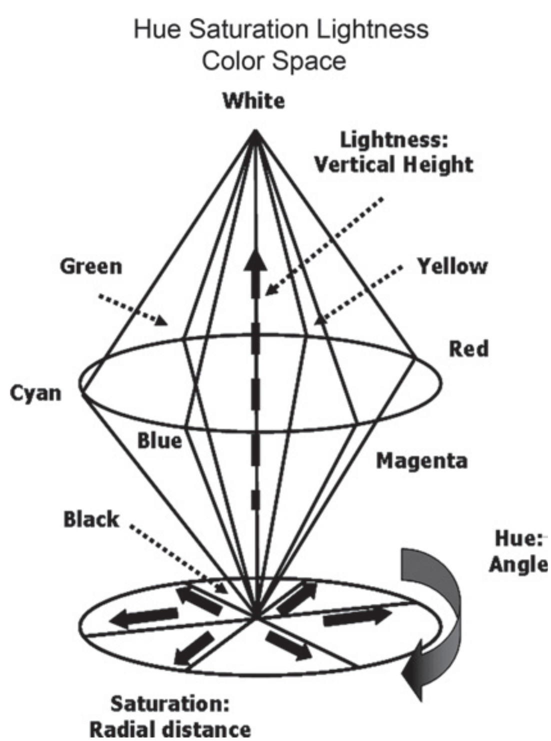


FIGURE 7.6 *The Hue Saturation Lightness Color Space. Similar to the HSV space except that the top consists of another conical shape with white at the very top.*

The dimensions in these two spaces are easily converted by simple linear equations to either the additive or subtractive color spaces—or both. In addition, there is a built-in constraint involving saturation and either value or lightness. In the HSV space, if the brightness of a highly saturated and very bright color is decreased, the saturation will have to decrease as well. In the HSL space, the same highly saturated and fairly bright (L = 0.5) color will lose saturation as it is made either brighter or darker. But this constraint is similar to how people see. In fact, both of these spaces behave more like a person's responses than either the additive or subtractive colors spaces.

These color spaces are in polar coordinates. The value or lightness axis is linear and runs up through the center of the space. But the saturation is given by the length of the radius vector, and the hue is given by the angle between some arbitrary starting point and the location of a given color.

CIE/Lab SPACE

In the 1930s, the Commission Internationale de L'Éclarige examined how people see colors, and developed the CIE color space. Its dimensions were denoted as x, y, and z and were based on a so-called *standard observer*. They based their work on a color-matching scheme in which many subjects were asked to examine small patches of color and arrange them as per their similarity or difference. This was a monumental step forward in that it gave a better idea of how people actually see colors. The result often is shown as the 1931 x,y chromaticity diagram. This had a few problems as well in that visually perceived differences among more widely spaced colors were different in different portions of the diagram. To correct this and a few other limitations, the work was revised in 1976 with the creation of the CIE/Lab color space. This is a three-dimensional object along the lines of the HSV space, but it is not as simple. One cut through the solid figure, perpendicular to the luminance axis, is shown in Figure 7.7. There is still a line that goes up through the center of the space, the L dimension, and it is representative of brightness. But the diagrams found in each of the planes that are perpendicular to this axis are not round. The overall object is more like an underinflated (American) football, with one side flattened. It comes to a point at the black end at the bottom, it has a noncircular shape that gets larger toward the more central values of L, and then comes back to a rounded point at the top, which is white. But the top and bottom halves are not the same in shape. The dimensions of the cross-sections of the football-like object are a and b. The three dimensions of the object are L, a, and b—hence the name, CIE/Lab—where a and b are in Cartesian coordinates in each plane intersected by the L axis.

Totally unlike the additive and subtractive color spaces, the primaries are derived mathematically and when considered as "colors" (that is, described spectrally) they wind up with some negative energy values at

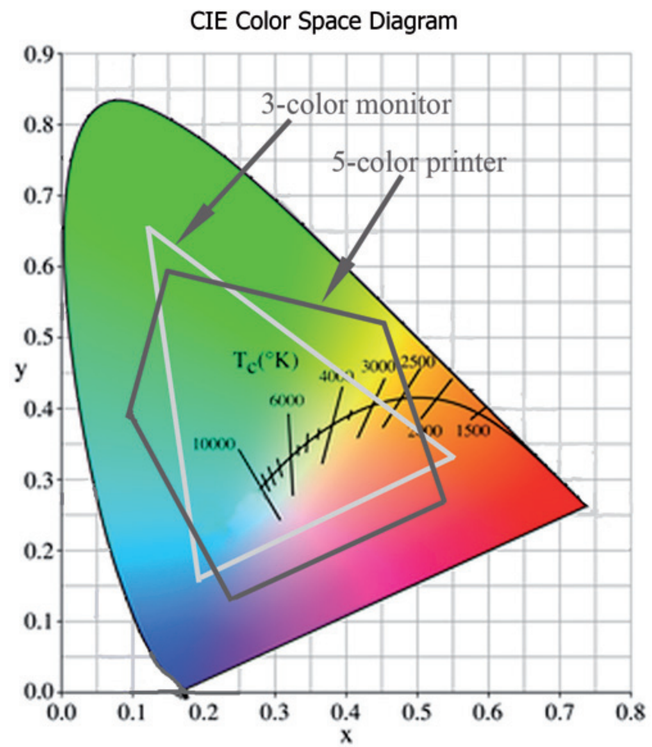


FIGURE 7.7 *The CIE Color Space Diagram. This was derived from how people see and its shape is suitably irregular. It is somewhat similar to the HSL space, but the color/saturation dimensions are mixed together in a Cartesian layout. It is not shown, but there is a vertical axis that goes below the page and rises above it, and this is the L, or luminance dimension. Two polygons are shown on the diagram: one for a color monitor and the other for a printer. It also shows the locus where "white" would plot for a number of color temperature rated sources.*

certain wavelengths. Clearly this is not possible as a real color, and so the primaries L, a, and b are considered to be theoretical. Also, the conversion from Lab to any of the previously described color space primaries involves nonlinear transformations.

Despite its complexity, the CIE/Lab space can be quite valuable. First of all, the space contains all the colors that humans can see. A point in the space can be thought of as a color of light. If a line is drawn connecting two such points, the line goes through all the colors that can be produced by adding together different amounts of those two lights. If a third light is added, three lines can be drawn to connect all three lights and the space within the triangle will contain all the colors that can be produced by those three lights. The color monitor triangle depicted in Figure 7.7 shows this. The number of colors that can be produced by a system is known as its *color gamut*. Shown in the figure is the gamut for a three-phosphor computer screen and a five-color printer.

Another important property of the CIE/Lab space is that a person's subjective impression of image sharpness is strongly determined by the L channel and only marginally by the other two channels. We can take some liberties with the spatial resolution in the a and b channels, but not in the L channel. Since the L channel is more strongly linked to the green channel in the RGB space (or the magenta channel in the CMY space), great care needs to be exercised with respect accordingly, these colors in their respective spaces are green and magenta. This has significant implications for image-processing modules. With efficient algorithms to convert from CIE/Lab to the other spaces and back, it is possible to take advantage of the visual system-emulating properties of the CIE/Lab space in making image calculations. Basically, the given image-processing tool will convert from the space chosen by the operator, most frequently RGB, into CIE/Lab. Then it will make the indicated calculations and convert the result back to the original space. There is an exercise on the website to demonstrate this effect, it is called, "L vs a, b channels".

YCC COLOR SPACE

This color space is similar to the HSL and the CIE/Lab in some respects. The Y channel, called the luminance channel, is a brightness dimension, much as the L channels in both the HSL and CIE/Lab spaces. The two C channels are somewhat similar to the a and b channels of the CIE/Lab space and have no intuitive relationship to the hue and saturation dimensions of HSL. The YCC space is linearly related to all the common color spaces except CIE/Lab. It was developed for use in broadcast television. Basically it allows for easy encoding of the color image information on the broadcast carrier wave and then subsequent decoding of the transmitted information to give back the color image. The result is compatible with black-and-white TV. The Y channel is essentially the black-and-white signal, and the other two channels are carried by a time-shifting process.

EXERCISES

Estimating Color System Response

In this chapter, it was shown that the response of a sensor element is given by the following formula:

$$R = \int_{\lambda} (a^{*}L)^{*}(b^{*}O)^{*}(c^{*}F)^{*}(e^{*}S)\, d\lambda$$

This exercise uses a spreadsheet processor to enable testing various combinations of elements in order to get a more intuitive understanding of the meaning of the relationships. It performs an approximation of the calculation indicated in the equation in that it multiplies the values for each wavelength and then sums the resulting products. The spreadsheet has a series of predetermined light sources (L), objects (O), filters (F), and sensors (S). Each of these is presented as a column of values at each of several wavelengths. For each experiment, choose the indicated elements and read the response in both linear and logarithmic units. Open the Color Estimator Spreadsheet that can be found on the web site.

There are five pages to the workbook; each is accessed via the tabs near the bottom of the screen. The first page is the Calculator. It accepts the various inputs and calculates the response values. It also shows two graphs. One is a bar graph that shows the relative contributions for each of the elements. The other graph is a smooth line graph that shows each element, wavelength by wavelength—that is, in "spectral" values. This is useful in understanding why a particular result is obtained.

The next four pages contain the elements. On the Sources page are spectral outputs of several light sources; the Filters page has spectral transmittance values for several filters. Some have fixed values and three have adjustable values. There is a sheet of the spectral reflectance values for a number of subjects—think of these as patches. Finally, there are spectral sensitivities for a number of sensor types.

The first step is to name the options for each of the elements. Go to one of the pages and study the graphical representations for each of the elements. Look to see which wavelengths are passed or reflected by filters and patches and which are blocked or absorbed, and figure out what that element option is. For example, a patch that reflects blue and green light but blocks red would be a cyan patch. A filter that passes blue and red light would be a magenta filter. Similarly, the light sources should be matched with typical types that have the spectra shown, and the sensors should be named after known types that have the indicated responses. This done, one can move on to conducting and analyzing experiments.

On the Calculator page, select the indicated element options, examine the graphical representation, and record comments regarding the graphs and the two response values—linear and logarithmic. Please note that the calculator multiplies the values across each wavelength to get the response at that wavelength, and then adds the responses at each wavelength. A few constants have been imbedded in the calculations to keep the graphical representations reasonable. Also, the calculator is fairly crude in that it works with wide wavelength bands—25 nanometers each. Another limitation is that the various input elements are only approximations of real filters, sources, sensors, and patches, but they give the idea of how the elements combine to produce results in real applications.

Experiments

By now you should have named all the options on each of the selections pages. Compare your results with those indicated in the Discussion section.

To simplify the setup of each experiment, the selections will be given for each combination to be evaluated. The organization is best shown by example as follows:

	Sensor	Filter #1	Filter #2	Object	Source	Linear Response	Logarithmic Response
1	Silicon	Yellow -.5	No Filter	Gray Card	Tungsten		

The idea is to make the indicated selections, read off the responses, and fill in the last two boxes.

For any given experiment, there will be several rows, where each row is one set-up. When all are done, compare the results and explain why the result came out the way it did. For example you might note that "the light source had a lot of red light, but the filter blocked most of that." You can compare your analyses with those in the Discussion section.

Experiment 1: Alternative Light Source Simulation

The computer has no way of dealing with fluorescence, so it is simulated in two parts. In the first part the light used to excite the sample is studied, then the light emitted is studied.

In the typical ALS setup, short wavelength light is introduced and a longer wavelength light is emitted. In the first four lines, daylight with a lot of blue is introduced, but the image is viewed through a red filter—first by a blue cone and then by a red one. The red object simulates the emitted light and the blue object indicates the background. In lines 5 through 10, a silicon chip is used instead of an eye. The yellow filter blocks the short wavelength light and the red object simulates the emitted light as before. In lines 5–7, a UV/blue light source is used. The red and green filter cases are there since red light plus green light gives yellow light, so the combination of these two should be similar to the yellow filter. The series is repeated with a daylight source, which adds contamination light.

	Sensor	Filter #1	Filter #2	Object	Source	Linear Response	Logarithmic Response
1	Blue Cone	Red	No Filter	Red	Daylight		
2	”	”	”	Blue	”		
3	Red Cone	”	”	Red	”		
4	”	”	”	Blue	”		
5	Silicon	Yellow -1	”	Red	UV		
6	”	Red -1	”	Red	UV		
7	”	Green -1	”	Red	UV		
8	Silicon	Yellow -1	”	Red	Daylight		
9	”	Red -1	”	Red	Daylight		
10	”	Green -1	”	Red	Daylight		

Experiment 2: Zero Lux (Night Vision)

Some cameras are called zero lux, which seems to imply that they can take pictures with no light. In fact the real meaning is that there is very little (almost zero) visible light, and the camera is working with infrared light.

	Sensor	Filter #1	Filter #2	Object	Source	Linear Response	Logarithmic Response
1	Silicon	IR	No Filter	Gray Card	Tungsten		
2	Rod	”	”	”	”		

The IR filter passes infrared and blocks almost all the visible portion of the spectrum (as typified by the response of a rod). The tungsten source produces a lot of IR light. The silicon sensor is quite responsive in the IR but the rod is a visual neutral, with very little IR sensitivity.

Experiment 3: Human Vision Compared to Silicon

Human color vision is accomplished by means of the three colored cones in the retina and is reasonably well balanced across the spectrum. Digital cameras, by comparison, use silicon sensors that have increasing sensitivity out into the near-IR.

The visual responses should be similar across the spectrum and the silicon responses should be systematically variable.

Experiment 4: Additive Lights and Subtractive Filters

The additive lights are red, green, and blue and their respective subtractive filters are cyan, magenta, and yellow.

	Sensor	Filter #1	Filter #2	Object	Source	Linear Response	Logarithmic Response
1	Silicon	Red	No Filter	Gray Card	Daylight		
2	Red Cone	”	”	”	”		
3	Silicon	Green	”	”	”		
4	Green Cone	”	”	”	”		
5	Silicon	Blue	”	”	”		
6	Blue Cone	”	”	”	”		

	Sensor	Filter #1	Filter #2	Object	Source	Linear Response	Logarithmic Response
1	Rod	Red	Cyan - 1	Gray Card	Daylight		
2	”	”	Cyan - 0.5	”	”		
3	”	Green	Magenta - 1	”	”		
4	”	”	Magenta - 0.5	”	”		
5	”	Blue	Yellow - 1	”	”		
6	”	”	Yellow - 0.5	”	”		

The use of a rod sensor gives some response across the spectrum, albeit weaker at the two ends. The filters are put in to be opposite each other and the strength of the subtractive filters is varied since they are normally used in varying amounts. The gray card and the daylight source tend to make these elements more or less uniform in this analysis.

Experiment 5: Color Adjustment

By using a combination of filters and light sources, this exercise will show how we can make a gray object appear to be brown.

Brown is made up of a lot of red light, significant green light (making orange), and a bit of blue light. The cyan filter tends to take daylight and make it close to this mixture. Using tungsten instead of daylight increases the amount of red relative to the green.

	Sensor	Filter #1	Filter #2	Object	Source	Linear Response	Logarithmic Response
1	Blue Cone	Cyan – 0.6	No Filter	Gray	Tungsten		
2	”	Gray	”	Brown	Daylight		
3	Green Cone	Cyan – 0.6	”	Gray	Tungsten		
4	”	Gray	”	Brown	Daylight		
5	Red Cone	Cyan – 0.6	”	Gray	Tungsten		
6	”	Gray	”	Brown	Daylight		

Experiment 6: Light Source Color Correction

Clearly if the light source has a color cast that is not what the camera expects, the pictures will not come out properly color-balanced. With photographic film, filters are used to correct for the light source, whereas with digital cameras it is possible to compensate electronically and mathematically. The digital camera solution has the effect of emulating the use of filters. In this exercise the use of two light sources and two filters will be used in parts 1 and 2.

Part 1: Using a Skylight Filter to Compensate for Daylight and Its Blue Cast

	Sensor	Filter #1	Filter #2	Object	Source	Linear Response	Logarithmic Response
1	Blue Cone	Skylight	No Filter	Gray	Daylight		
2	”	”	”	”	Tungsten		
3	Green Cone	”	”	”	Daylight		
4	”	”	”	”	Tungsten		
5	Red Cone	”	”	”	Daylight		
6	”	”	”	”	Tungsten		

The three color sensors in the eye are exposed to both daylight and tungsten filters through a filter that is designed to compensate for the blue cast of daylight. The object is a gray card and should evoke nearly the same response for each color sensor.

Part 2: Using a Tungsten Control Filter to Compensate for the Reddish-Yellow Cast of Tungsten Light

This part is very similar to the previous one, the only changes being the filter and the light source.

	Sensor	Filter #1	Filter #2	Object	Source	Linear Response	Logarithmic Response
1	Blue Cone	Tungsten Control	No Filter	Gray	Daylight		
2	"	"	"	"	Tungsten		
3	Green Cone	"	"	"	Daylight		
4	"	"	"	"	Tungsten		
5	Red Cone	"	"	"	Daylight		
6	"	"	"	"	Tungsten		

Experiment 7: Filter Factor

Whenever a filter is introduced, it blocks light and thereby requires an increase in the camera's exposure settings to compensate. There are two components to this: the increased absorption in certain parts of the spectrum and the effects of filter material (often glass) in terms of front surface reflection. This component is small (usually less than 0.04 log exposure) and it is shown explicitly in this exercise.

Rows 1 and 2 give the responses with no filters, and rows 3 and 4 give the responses with the relatively clear skylight filter in place. Comparisons should be made, with and without, for the same light source conditions.

	Sensor	Filter #1	Filter #2	Object	Source	Linear Response	Logarithmic Response
1	Visual Neutral	No Filter	No Filter	Gray	Tungsten		
2	"	"	"	"	Daylight		
3	"	Skylight	"	"	Tungsten		
4	"	"	"	"	Daylight		

White Balance

White balance is the process of removing color casts from the image so that the colors of the original scene, as viewed by a person, are rendered the same color in a photo.

Correct camera white balance has to take into account the "color temperature" of a light source. The eyes are very good at judging what is white under different light sources; digital cameras, however, often have great difficulty with auto white balance (AWB).

An incorrect WB in the camera can create color casts, which are unrealistic and particularly damaging to crime scene photographs.

As you have already studied, the color of each light source is categorized by its color temperature in degrees Kelvin (K).

The table in Figure 7.E1 is a rule-of-thumb guide to the correlated color temperature of some common light sources.

All digital cameras have controls for setting the color balance. The amount of controls vary from camera to camera, and the descriptions and symbols may vary, but they have the same functions. The descriptions are just rough estimates for the actual lighting under which they work best.

Color Temperature	Light Source
1000-2000 K	Candlelight
2500-3500 K	Tungsten Bulb (household variety)
3000-4000 K	Sunrise/Sunset (clear sky)
4000-5000 K	Fluorescent Lamps
5000-5500 K	Electronic Flash
5000-6500 K	Daylight with Clear Sky (sun overhead)
6500-8000 K	Moderately Overcast Sky
9000-10000 K	Shade or Heavily Overcast Sky

FIGURE 7.E1

The table in Figure 7.E2 is an example of some controls found on a digital camera. The first three white balances allow for a range of color temperatures.

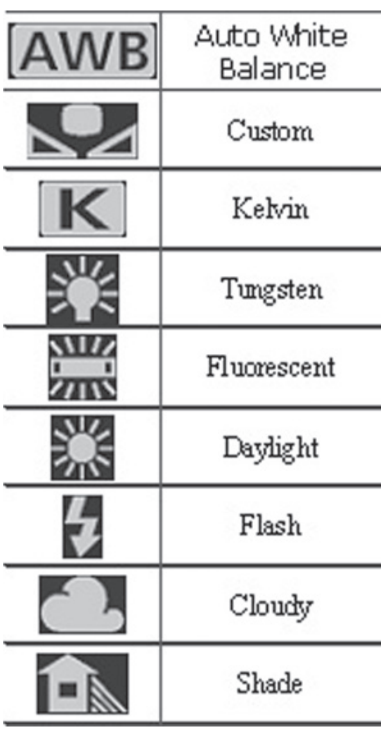


FIGURE 7.E2

Auto white balance is available in all digital cameras and uses a best-guess algorithm within a limited range—usually between 3000/4000 K and 7000 K.

Custom white balance allows you to take a picture of a known gray reference under the same lighting, and then set that as the white balance for future photos.

Kelvin allows you to set the color temperature over a broad range. This is very useful if you know the color temperature of the light source.

The remaining six white balances are listed in order of increasing color temperature; however it will vary from camera to camera.

Some subjects can create problems for a digital camera's auto white balance—even under normal daylight conditions. One example is if the image already has an overabundance of warmth or coolness due to unique subject matter. An example might be a situation where the subject is predominantly red—the camera mistakes this for a color cast induced by a warm light source. The camera then tries to compensate for this so that the average color of the image is closer to neutral, but in doing so it unknowingly creates a bluish color cast on other objects in the image. Some digital cameras are more susceptible to this than others. A digital camera's auto white balance is always more effective when the photo contains at least one white or bright colorless object.

It should be mentioned here that there are those who say the AWB setting is the last option you should use, no matter what. Others contend that to get the best color balance, you must shoot in the RAW format and adjust the color balance after the image has been acquired.

The question arises as to where the WB on a camera should be set. The best answer possible is to run some tests using different settings and lighting. Select what gives you the best results.

Channels

There are basically three mode settings (channels) in PhotoShop:

- RGB
- CMYK
- Lab

Channels are grayscale images that store different types of information:

Color information channels are created automatically when you open a new image. The image's color mode determines the number of color channels created.

Alpha channels store selections as grayscale images. You can add alpha channels to create and store masks, which let you manipulate or protect parts of an image.

Spot color channels specify additional plates for printing with spot color inks.

An image can have up to 56 channels.

By default:

- Bitmap, grayscale, duotone, and indexed-color images have one channel.
- RGB and Lab images have three.
- CMYK images have four.
- You can add channels to all image types except Bitmap mode images.

Note: As long as you save a file in a format supporting the image's color mode, the color channels are preserved. Alpha channels are preserved only when you save a file in Photoshop, PDF, PICT, Pixar, TIFF, or raw formats.

One of the many uses of channels when working with RGB images is the capability of individually enhancing the information in each channel.

Understand that you are adjusting the color information in that channel and the color balance will shift. Figure 7.E3 shows how the channels can be found. This technique is useful only if the color balance is not important and you are just interested in enhancing the detail, such as in a fingerprint. Study your image carefully before using this technique; you don't want to lose any information.

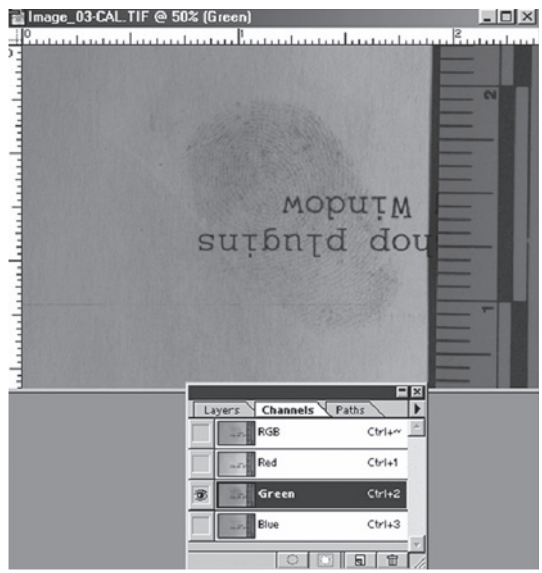


FIGURE 7.E3

1 Open Image_03.
2 Click on each channel to select it. Study the information in each channel. The red channel has no information on the fingerprint.
3 Carefully enhance the blue and green channel using the curves tool.
4 Select the RGB composite channel to see the results of the enhancements.
5 Convert the image to grayscale and compare the original with your enhancement as shown in Figure 7.E4.

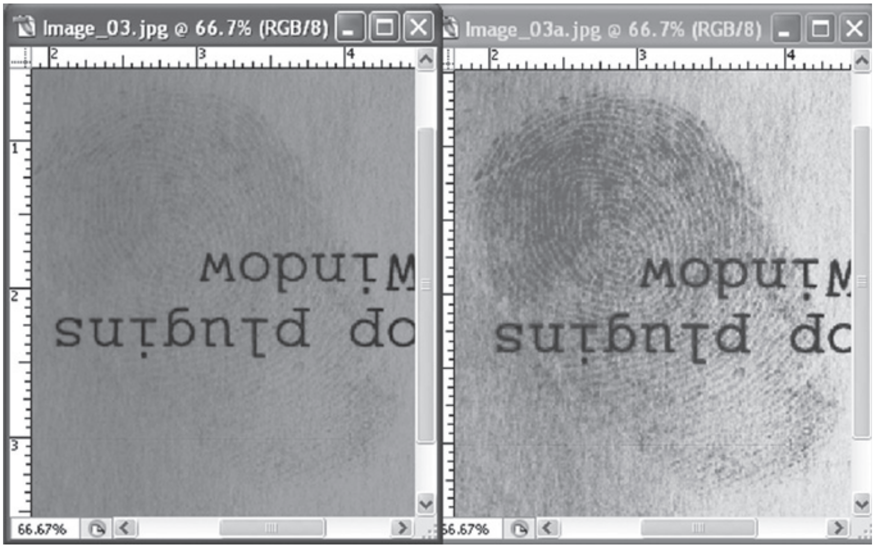


FIGURE 7.E4

Using Calculations with Channels

Many times you can use the calculations command to blend two channels together; thereby enhancing the image.

This involves using the correct blending mode in PhotoShop. It is necessary to understand what each blending mode does to the image. To understand the description of each mode, note the following:

- The base color is the original color in the image.
- The blend color is the color being applied with the painting or editing tool.
- The result color is the color resulting from the blend.

Blending Modes

Normal

- Edits or paints each pixel to make it the result color.
- This is the default mode. (Normal mode is called Threshold when you're working with a bitmapped or indexed-color image.)

Overlay

- Multiplies or screens the colors, depending on the base color.
- Patterns or colors appear to lay on top of the existing pixels while preserving the highlights and shadows of the base color.
- The base color is not replaced but is mixed with the blend color to reflect the lightness or darkness of the original color.

Difference

- Looks at the color information in each channel and subtracts either the blend color from the base color or the base color from the blend color, depending on which has the greater brightness value.
- Blending with white inverts the base color values; blending with black produces no change.

Screen

- Looks at each channel's color information and multiplies the inverse of the blend and base colors.
- The result color is always a lighter color.
- Screening with black leaves the color unchanged.
- Screening with white produces white.
- The effect is similar to projecting multiple photographic slides on top of each other.

Multiply

- Looks at the color information in each channel and multiplies the base color by the blend color.
- The result color is always a darker color.
- Multiplying any color with black produces black. Multiplying any color with white leaves the color unchanged.
- When you're painting with a color other than black or white, successive strokes with a painting tool produce progressively darker colors. The effect is similar to drawing on the image with multiple magic markers.

You can use the blending effects associated with layers to combine channels within and between images into new images using the Apply Image command (on single and composite channels) and the Calculations command (on single channels).

These commands offer two additional blending modes not available in the Layers palette—Add and Subtract. Although it's possible to create new combinations of channels by copying channels into layers in the Layers palette, you may find it quicker to use the calculation commands to blend channel information.

The calculation commands perform mathematical operations on the corresponding pixels of two channels (the pixels with identical locations on the image) and then combine the results in a single channel.

Two concepts are fundamental to understanding how the calculation commands work; refer to Figure 7.E5:

FIGURE 7.E5

- Each pixel in a channel has a brightness value from 0 ("off" or black) to 255 ("on" or white). The Calculations and Apply Image commands manipulate these values to produce the resulting composite pixels.
- These commands overlay the pixels in two or more channels. Thus, the images used for calculations must have the same pixel dimensions.

1 Open Image_02.
2 Crop the image to just the fingerprint.
3 Study the information in each channel in the channels palette. Note that all the information in the fingerprint is in the red and green channels. The blue channel has no information at all.

This step determines which channels will be blended together. If there is information in all channels, do not use this technique, because you will be discarding information from the image. Refer to Figure 7.E6.

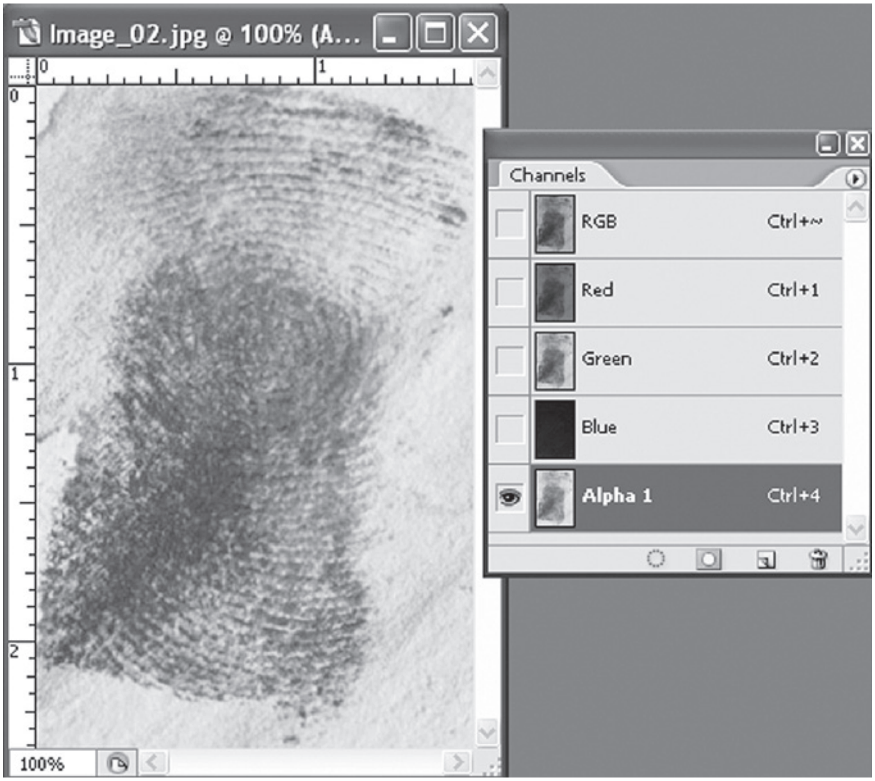


FIGURE 7.E6

1 Select the **Image menu** > **Calculations command**.
2 Set the source 1 channel to red.
3 Set the source 2 channel to green.

Note these are the two channels that contained all the information.

4 Set the blending mode to screen.
   - The screen mode is used, because it looks at each channel's color information and multiplies the inverse of the blend and base colors.
   - The result color is always a lighter color.
5 Click **OK** to apply the calculations.

An Alpha channel is created in the channels palette by the new image as shown in the figure to the right.

Alpha Channels and Masks

When you select part of an image, the area that is not selected is "masked," or protected from editing. So when you create a mask, you isolate and protect areas of an image as you apply color changes, filters, or other effects to the rest of the image. You can also use masks for complex image editing, such as gradually applying color or filter effects to an image.

Examples of Masks

Refer to Figure 7.E7:

A Opaque mask as used to protect the background and edit the object
B Opaque mask as used to protect the object and color the background
C Semitransparent mask as used to color the background and part of the object

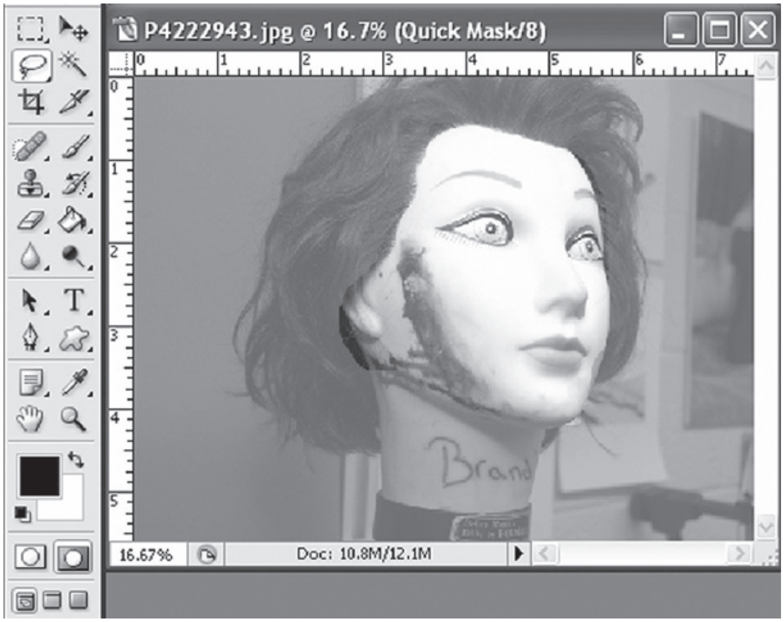


FIGURE 7.E7

In Photoshop, masks are stored in Alpha channels.

- Masks and channels are grayscale images, so you can edit them like any other image.
- With masks and channels, areas painted black are protected, and areas painted white are editable.

Quick Mask Mode

This function lets you edit any selection as a mask. The advantage of editing your selection as a mask is that you can use almost any Photoshop tool or filter to modify the mask.

1 Ensure that the default colors are set in PhotoShop (black foreground, white background).
2 Open the image of Brandi, **P4222943.jpg**.
3 Create a selection with the Lasso tool around Brandi's face.
4 Select the Quick-Mask mode tool at the bottom of the tool box. A red mask appears around the selection.
5 The paintbrush tool can be used to expand or decrease the selection. Select the paintbrush tool and set the size.
6 To subtract from the selection:
   - Set the foreground color to black.
   - Paint over the area that you want to remove from the selection.
7 To add to the selection:
   - Set the foreground color to white.
   - Paint over the area that you want to add to the selection.
8 Turn the Quick-mask mode from off to on to see the actual selection marquee.
9 Continue until you have an accurate selection marquee around the face.

Selections you make using the Quick Mask mode can be saved and loaded using Alpha channels.

Alpha Channels

Alpha channels can be used to save and load selections. Refer to Figure 7.E8. You can edit Alpha channels using any of the editing tools. When a channel is selected in the Channels palette, foreground and background colors appear as grayscale values. Storing selections as Alpha channels creates

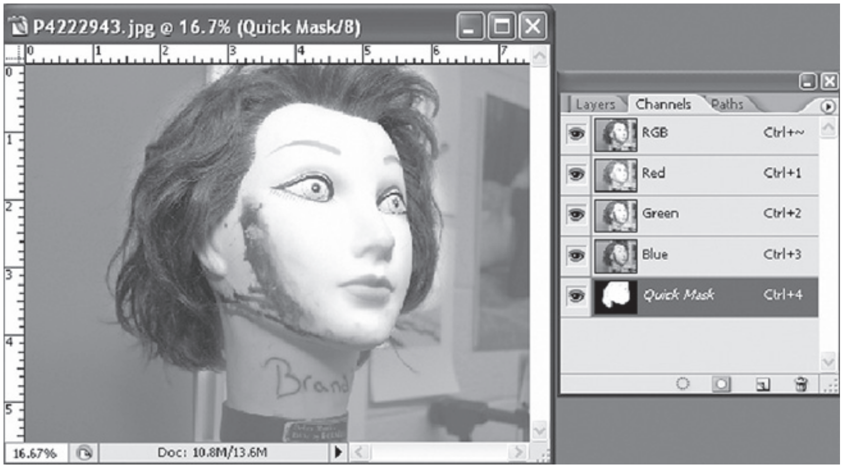


FIGURE 7.E8

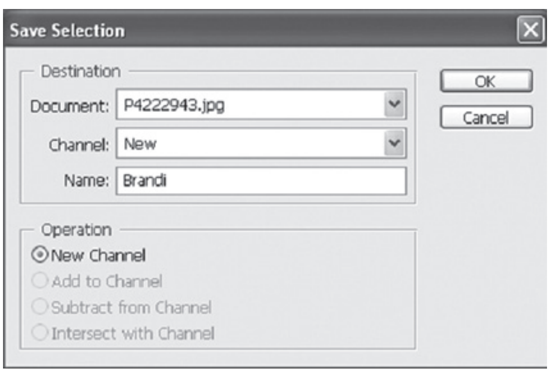


FIGURE 7.E9

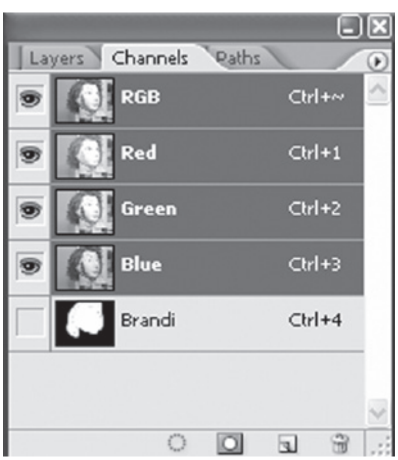


FIGURE 7.E10

more permanent masks than the temporary masks of Quick Mask mode.

You can reuse stored selections or even load them into another image. The mask appears as an Alpha channel in the channels palette as shown to the right.

Saving

To save a selection as an Alpha channel, see Figures 7.E9 and 7.E10:

1 Turn off the Quick-mask mode and a selection marquee appears in the image.
2 Access the **Select** > **Save Selection menu**.
3 Name the selection.
4 Click **OK**.

The selection now appears at the bottom of the channels palette as an Alpha channel. Note that when you save the image, Alpha channels can only be retained in Photoshop, PDF, PICT, Pixar, TIFF, or raw formats.

Lab Color Mode

The Lab Color mode has a lightness component (L) that can range from 0 to 100. In the Adobe Color Picker, the a component (green-red axis) and the b component (blue-yellow axis) can range from +127 to −128.

The Adobe Color Picker lets you select a color based on the Lab color model.

- The L value specifies the luminance of a color.
- The A value specifies how red or green a color is.
- The B value specifies how blue or yellow a color is.

You can use Lab mode to work with Photo CD images, edit the luminance and the color values in an image independently, move images between systems, and print to PostScript Level 2 and Level 3 printers. To print Lab images to other color PostScript devices, convert the image to CMYK first.

Lab images can be saved in Photoshop, Photoshop EPS, Large Document Format (PSB), Photoshop PDF, Photoshop Raw, TIFF, Photoshop DCS 1.0, or Photoshop DCS 2.0 formats.

You can save 48.bit (16 bits per channel) Lab images in Photoshop, Large Document Format (PSB), Photoshop PDF, Photoshop Raw, or TIFF formats.

Note: The DCS 1.0 and DCS 2.0 formats convert the file to CMYK when opened.

Note: Lab color is the intermediate color model Photoshop uses when converting from one color mode to another.

Chapter 8

Showing Images

Images come in two distinct types: hard copy and dynamic (self-illuminating). Hard copy images are called prints and can be either reflective prints, such as normal snapshots, or they can be transparencies, such as slides. To make reflective prints, a colorant—either an ink or a dye—is deposited on a sheet of paper. In the case of transparencies, the colorant is deposited on a transparent base. Dynamic images are displayed on a screen by some active device, such as a TV or a computer projector. Traditional movies are projected onto a screen but exist on a strip of film. The film-resident images are prints and the screen-resident images are dynamic.

FRONT SURFACE REFLECTION

Let's say you just received a letter from your mother. It is in the old style—ink on paper—and delivered by the postal carrier. You hold up the letter and can easily see the characters on the page. Then you hold it behind a sheet of glass. You can still see it clearly except in a few areas. In those areas, the light from the lamp over your shoulder is reflecting from the sheet of glass. Now let's imagine that you had picked up a piece of frosted glass instead of a sheet of clear glass. You can now see only the larger characters on the page; and what was black is now sort of gray. Finally, let's say you pick up a mirror. Now you cannot see the letter at all. No matter how much ink there is on the paper, all you see is the light from behind you reflecting off the mirror.

We have just examined the concepts of front surface reflection and a close cousin, scattering. The glass, the frosted glass, and the mirror all are reflecting some of the incoming light back to your eyes. The clear glass reflects the least, and the mirror reflects the most. The frosted glass, in addition to reflecting some of the light back to you, is reflecting bits of the light at random angles. In all cases, front surface reflection is preventing you from having a clear view of the ink below on the letter.

When light goes through one medium, such as air, and then encounters another medium with a different index of refraction, such as glass, plastic, or gelatin (as in photographic film and paper), a portion of the incoming light will be reflected back off the front surface. The index of refraction is the ratio of the speed of light in the given medium divided into the speed of light in a vacuum (air, in this circumstance, might as well be a vacuum). Light travels more slowly in a material with a high refractive index. The bigger the shift in refractive index, the more the light will be reflected off the front surface. *Some amount of this reflection is unavoidable*. Camera lenses suffer from front surface reflection. Special coatings are applied to minimize the effect, but it cannot be eliminated. By the way, if the light were coming from inside the material with the high refractive index and arriving at the interface with the low refractive index, the same sort of reflection would take place. Some of the light would bounce back into the glass instead of exiting. This is called internal reflection; and if the angle is steep enough, the result is total internal reflection. Fiber optics depend on total internal reflection to keep the light inside the conductor.

Consider what you see when you watch television. There are bright areas and dark areas in the picture. Some are such that your mind calls them white and others that you consider black. But even though a television screen can get rather bright in some areas, it cannot get any darker than that same screen when the set is turned off. Yes, the gray screen of the turned-off TV is the black you think you are seeing when you watch the turned-on set. Behind the glass front of the screen there is a black shadow mask to help with colors and, to some extent, reduce glare. But what reflects off the front surface is totally unaffected by what is on the other side of the glass. So why is the screen gray when it is not turned on? Because light is reflecting off the front surface of the glass! It does not matter how black the shadow mask is, because light coming off the front surface of the glass never gets to the shadow mask. The reflected light comes from light in the room, reflects off the front of the glass, and goes into your eye. If you turn off the lights in the room, the blacks get blacker since there is very little light impinging on the front surface. And so while the *percent* reflecting remains the same, the *amount* reflecting is much smaller.

Front surface reflection is unavoidable. However the degree to which it causes problems can be controlled. When images are dynamically projected or shown on a screen, the room lights can be turned down. Showing transparencies by projecting them also calls for a darkened room. With reflection prints, higher room light is better, because the image has no light of its own. But it should normally be diffuse illumination—seemingly coming equally from all directions at once. Specula light (seemingly traveling in parallel beams from a certain light source) can be used if it is from only a few sources and those beams impinge at an angle of about 45 degrees to the viewing angle. In this way, the glare off the surface is reduced.

Some of the papers used to make photographic prints have a roughened surface, called a matte surface. Matte surfaces are designed to minimize the distraction caused by fingerprints on glossy surfaces. But the darker colors on matte-surfaced prints are not quite as dark as those on glossy-surfaced prints. Matte surfaces tend to give a "softer" look to the image, so they are sometimes used for sentimental portraits. Consider the application of patches of paint to a test surface. We will select four different paints in two pairs. One pair is navy blue and the other is white. One of the paints in each pair dries to a glossy surface and the other to a flat surface. The whites will appear to have virtually the same color, but the flat blue will not be as dark as the glossy blue (unless extra pigment was added intentionally to the paint to compensate). The flat blue appears to be slightly tinged with white (consider again the frosted glass on top of the letter from your mother).

The lesson to be learned is that front surface reflection will limit the maximum darkness we can achieve in a picture. If the display has its own light—as in a television, movie projector/screen, or slide projector/screen—we can turn off the room lights and the darks can be made darker, since there is very little stray light to reflect off the surface. But with hard copies—such as photographic prints, where turning off the room lights will make the whites go black as well as the blacks—there is little that can be done to ameliorate the dilution of the high densities. Also, flat or matte surfaces exaggerate the effect. As a rule of thumb, it is very difficult to get front surface reflections down below 0.5%. This equates to a maximum optical density of 2.3 or 1/200 of the incoming light. So even though a camera using portrait film may be capable of a dynamic range of 20,000:1, the print will be limited to a dynamic range of 200:1 or less. There are several ways to get all the important information from a camera image onto a print. Some techniques include reduction of overall contrast; putting a curvature into the response curve; increasing the brightness in certain portions of the image or decreasing the brightness in other areas (known as dodging and burning); or, if there are several shots of the same scene at different exposure levels, merging those images. So even if a print cannot represent all the information in the original, it is important that the information is there to be enhanced and represented in the print.

To measure the optical reflection density (log of the amount of incoming light that is absorbed—*absorbance*) of a photographic image, we can use a device called a densitometer. Because of the issues raised so far, the device requires a particular geometry in order to work properly. This is called the *45–90 geometry*. The device illuminates the spot to be read with a specula beam of light coming in at 45 degrees. It measures the reflected light with a specula beam rising off the surface at 90 degrees. The bulk of the front surface reflection will exit the spot being read at 45 degrees and only a small fraction will come off at 90 degrees. The typical front surface reflection limit value of 0.5 degrees is read in this way. The copy stands used in laboratories to photograph items of interest are also set up to facilitate 45–90 geometry.

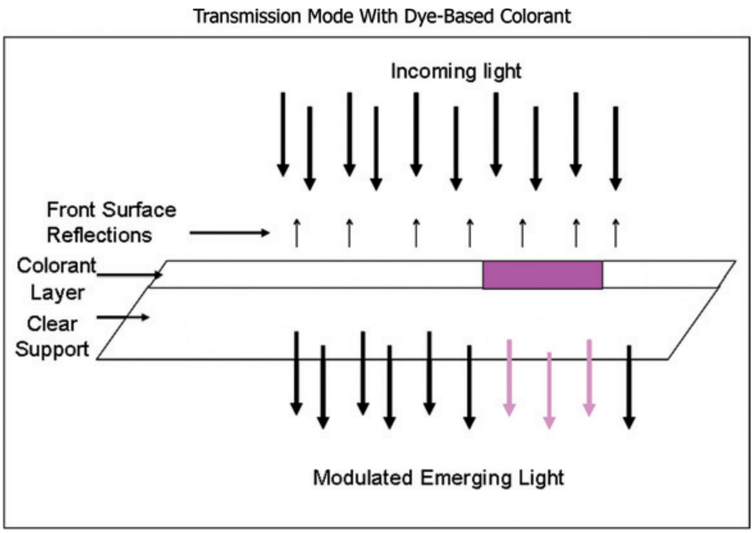


FIGURE 8.1 *Transmission Mode. Schematically the process by which light goes through a transmission sample is depicted. Some light is reflected off the front surface and the rest goes through either a clear portion or a colored portion. The light goes through the clear potion without change and light that is not absorbed by the dye goes through the colored portion.*

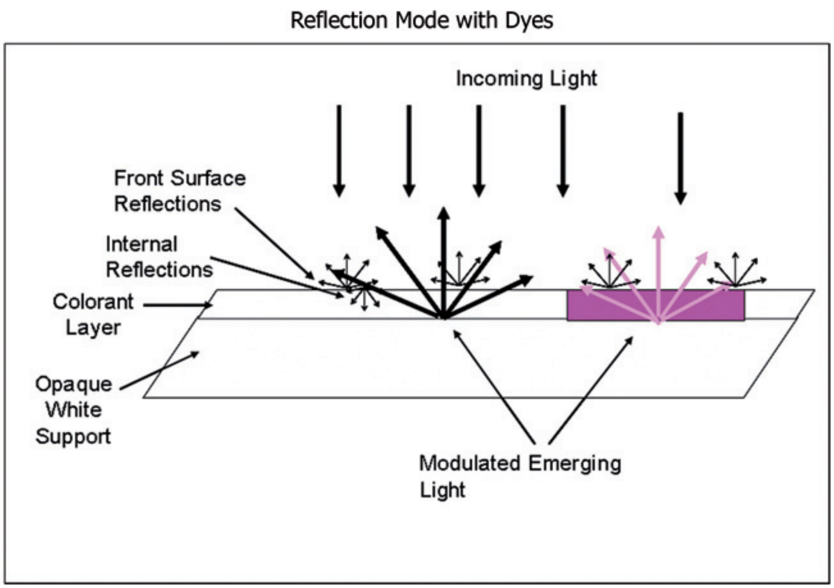


FIGURE 8.2 *Reflection Mode. Schematically the process by which light responds to a reflection sample is depicted. Some light is reflected off the front surface. The rest goes either through a clear portion or a colored portion. The light goes through the clear portion, reflects off the white substrate, and most of it comes back out the top as white. Some of it reflects off the layer's top surface. The portion that goes through the color patch behaves in a similar way except that some of the light was absorbed by the dye. The front surface reflection limits how dark the colored patch can appear.*

The details of light going through a transmission material, such as a slide, are shown in Figure 8.1. Similarly the details of viewing a reflection print are shown in Figure 8.2. Note that the transmission mode is rather simple compared to the reflection mode. The bottom line is that in reflection mode there may be low density limitations and there will certainly be high density limitations.

Inks and dyes are used to make hard copy prints—either transmission (slides) or reflection (paper-based). The overwhelming majority are reflection prints. All reflection prints are limited by front surface reflections and the nature of the colorant will have some effect on this.

BASIC COLORANTS

In the days of Rembrandt, paints were colored with mineral pigments. These were made from certain chemical compounds. For example, titanium oxides give a very neutral white, various cadmium compounds give yellows, and cobalt compounds give blue. Also, some colorants were derived from biological sources. Some blues and purples were derived from extracts of the bodies of snails. Some reds are from berries. The list is long and many of the actual recipes were lost long ago. (For example, the blue dye used in the Jewish prayer shawl is supposed to be made according to a formula based on a certain sea creature. However, for the past 1000 years or so, no one has known just which sea creature that is supposed to be.)

Dyes

A dye is an agglomeration of a large number of special molecules in a transparent material. Typically those used in making images have a benzene ring with a tail attached at one (or more) of the six apex points. The make-up of the tail generally dictates which wavelength(s) of light the molecule will absorb. These are shown schematically in the upper part of Figure 8.3. In the case of photographic films and papers, the medium is clear gelatin. In dishwashing liquids, it is the water in the liquid detergent. In some filters, it is a plastic. And in stained glass, it is the glass itself.

The color of a suspension of dye molecules is determined by the wavelengths that are absorbed. If a dye absorbs blue light, it will appear yellow. That means it is absorbing blue and letting red and green pass unaffected. If the dye absorbs green light, it will appear magenta. It is blocking green light and letting blue and red light pass. A blue filter blocks yellow (that is, red and green) and lets blue light pass. The color of the filter is the color of the light that it lets pass and the opposite of the color it absorbs.

The suspension is very close to being a solution. In any event, the dye molecules have no discernable, external shape. The light-absorbing material appears simply to be dispersed in the medium. Since it has no external shape, it has no front surface and, therefore, no front surface reflection occurs when light approaches the dye molecule. It simply absorbs some of the photons at certain wavelengths and does not affect any other photons. The suspension appears colored but otherwise clear in that the light that does go through, goes straight through. We can clearly see right through the suspension at

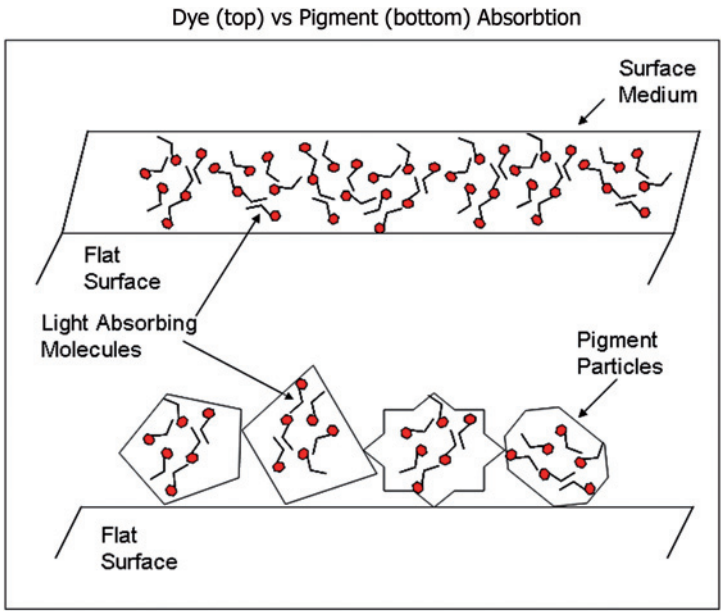


FIGURE 8.3 *Dyes vs. Pigments. Dyes are shown in the top image dispersed in a clear film. The dye does not cause scattering of incoming light. The lower image shows pigment particles on a surface. There will be scattering off the various surfaces of the particles, causing the color to be diluted with white light and preventing the placement of a different layer of pigment under the top layer.*

the nonabsorbed wavelengths. The only front surface reflection effects occur when the incoming light reaches the boundary of the total suspension. No such effects occur when the light reaches the dye molecule.

Dyes can be mixed together or applied in layers or dye can be superimposed. One can put two different dye molecules into the same suspension and there is no interaction. They each act as if they were there alone. Put a cyan and a yellow dye in the same gelatin layer and it will appear green. That is to say, red and blue light is blocked and green is allowed to pass.

If the dye is dispersed in a plastic and coated on a clear base, we can make a dye-based filter. If a cyan dye is put into one filter and a yellow dye is put into another, the two filters can be put into the same beam of light and the effects will combine just as if they were in the same layer. The cyan filter will pass green and blue light and the yellow filter will pass red and green light. Together they will pass only green and block red and blue. The result is essentially a green filter. To render grass in a photographic paper print, it is accomplished with dyes in the cyan and yellow layers, leaving the green to come through. The same is true with a dye sublimation printer.

The important factor is that dyes do not scatter light. They appear clear and absorb only some of the incoming light at the selected wavelengths. Accordingly they can be mixed together and superimposed.

One downside to dyes is that they fade over time. The effects of light, heat, and humidity accelerate the fading. Special chemistry is employed to make them more stable, but as it says on the warning labels, "in time, all dyes will fade."

Inks

The walls of the ancient Egyptian burial chambers still show vibrant colors, but the colors in the photos of you as a young child are not what they used to be—especially if the picture was hanging on a wall in sunlight. The ancients used pigments, or inks, whereas your photos are made with dyes. Inks are less susceptible to fading.

To make an ink, the dye is attached to a more stable, solid molecule. Take the dye used to make Easter eggs. It starts out as a dye, but the eggs—with their carbonate shells—are treated with vinegar (acid) and then the dye is applied. The dye attaches to the partially dissolved carbonate material (a mordent process) to produce a new combination that is more stable than the dye alone. Inks are made by attaching dye molecules to more stable materials in order to make the colorants more stable. Ink particles are shown in the lower part of Figure 8.1. This has certain side effects, however. The colorants are no longer as clear as the original dyes.

Consider the reverse process. In a hobby shop, you can find paints to be used on pottery. The label has to show what color the material will come out to be after it is fired in a kiln. Looking at the slurry in the bottle will give only a vague idea of what the ultimate color will be. The material appears chalky, grayish, and dense. But after the material is melted in the kiln and a layer of the colorant is dispersed in the glaze, it will be much more vibrant. If we were to take the filter discussed earlier (dye in plastic) and grind up the plastic, it would lose its vibrant color and start to approach a faintly colored grayish powder of the potter's paint. If the powder were to be melted down and allowed to cool into a smooth layer again, the vibrant color would return. The point is that inks are bound in rough, randomly shaped particles like sugar crystals, and they scatter light of all wavelengths before that light gets to the imbedded dyes. Inks will absorb some of the photons of the selected color of those landing on them, but they will also scatter light of all wavelengths. Some of the scattered light will go back to the observer. Very little light will actually penetrate the layer of ink (as was the case in a dye layer).

Because of the scattering, it is not feasible to put one ink on top of another as was done with the dyes. If a yellow ink were deposited on top of a cyan ink, the resulting deposit will essentially look yellow and only faintly green. This is because the yellow layer will absorb blue and scatter all colors. The scattering prevents much light from penetrating down to the cyan layer. It also prevents any light that had been modulated by the cyan layer from coming back through the yellow layer (it scatters on the way back as well). To make a green spot using inks, one would put down a very small spot of yellow ink and right next to it, a very small spot of cyan ink. One spot will provide blue absorption and the other will, in very close proximity, provide red absorption. If the result is too fine for the eye to resolve, the effects will be visually perceived as green. The scattering nature of inks also makes them marginal for transparencies. The projected images appear blurry. This pointillism effect is well known to magazine printers, certain impressionist painters, and computer screen manufacturers.

Remember that inks comprise small particles of material with imbedded dyes. Front surface reflection of light landing on the particles from any light that enters the patch is what causes the scatter. It is much harder to get vibrant colors from inks than from dyes.

THE MIDDLE GROUND

In one part of the previous discussion, a dye was dispersed in a clear plastic and then ground up to make an ink. The advantage was the possible increase in the stability of the colorant and the disadvantage was its scattering nature. Now let's consider coating the particles with a partially cross-linked polymer and dispersing the whole combination in a material that makes the polymer semi-soluble. When the material is deposited and the solvent starts to evaporate, the polymer will coat the ink particles and create something close to a smooth layer, with close to only one index of refraction. The polymer will continue to cross-link and become durable. The result is that the scattering from the individual particles will be greatly reduced and the layer will act more like dye in a clear medium. This produces a colorant system that has greater stability and more vibrant color capability.

The only disadvantage is a reduction in shelf life of the ink. Little bits of dye-impregnated glass dispersed in water will have a long shelf life. We might have to shake or stir the mixture before using, but the particles will disperse again as per the original state. Not so if there is a sticky polymer present. As the particles settle out, the polymer will begin to stick the particles together, and no amount of shaking or stirring will reverse the process completely. These improved inks tend to have limited shelf life.

It is difficult to make this type of ink. The optics, chemistry, and rheology must all be just right, but progress is being made. In addition to better color stability and more vibrant colors, another special result is better resolution. In dye sublimation printers, the size of an individual pixel is controlled by the size of the heating element in the printer. The smaller it is, the sharper the image can be. With such a printer, the three colors can be superimposed and the size of the pixel on the print is the same as the size of the drop of dye that is deposited.

With ink printers, the story is much more complex. The ink can be laid down in very small drops, but it takes several drops to make a pixel. Since the ink is essentially opaque, one drop of ink cannot be placed on top of another. They have to be side by side. Each drop blocks some of the white of the underlying paper surface.

In the simplest model of a black-and-white laser printer, portions of the paper are reserved for pixels. To make white, no ink is deposited in the reserved pixel area. To make black, the whole area is covered with dots of black ink. To get a level of gray that is just above white, one dot is deposited. The next level of gray is accomplished by depositing two dots, and so on, until the whole area is covered with dots. The total number of dots that is required will be equal to the bit depth of the image (let's say 256). If the array of dots is square then it is easy to see that the sides of the array will each be composed of 16 dots. The pixel will be composed of 16 rows and 16 columns of dots—16 times 16 is 256. That is, the pixel will have a number of dots on each side that is equal to the square root of the bit depth. This means that a printer rated at 1600 dots per inch, in the simplest case, will have an actual resolution of 1600 divided by the square root of 256 (16) and have a real resolution of 100 pixels per inch. Note that the dye sublimation printer may have a dot size that is 1/300 inch in size, which means it can lay down 300 dots per inch. But since each dot of a dye printer is actually a pixel, this printer has a resolution of 300 pixels per inch. In reality, in modern inkjet printers, there are ways to mathematically improve the resolution of the simple model shown here. But in any event, do not confuse dots per inch with pixels per inch. In dye sublimation printers, the number of dots per inch is the same as the number of pixels per inch. In ink and laser printers, the number of dots per inch is considerably higher than the number of pixels per inch. And since the viewer will be seeing pixels, the actual resolution will be lower than the dots-per-inch value.

As progress is made on new inks—ones that have more film-forming properties that emulate dye layers—it will be possible to approach the *one dot equals one pixel* formula of the dye printers. At the same time, there is also progress in printer designs. They are using more than just the CMYK colorants, which can reduce the needed pixel depth for each color. These also have special patterns to maximize coverage, and they change the bit depth at sharp edges. The result will be greatly enhanced print resolution and better image brilliance.

Finally it is important to remember that modern digital printers perform significant mathematical adjustments to the image sent to them by the computer. Resolution will be recomputed in order to fit within the parameters of the physical printing mechanism. Bit depth might be altered in certain portions of the image to enhance apparent sharpness. Tone scales might be adjusted to compensate for the reflectance properties of the materials used in the printer. All these adjustments are applied to all incoming image files and the adjustments are made to the entire image, not just preselected parts. They will not alter the image in any way that will lead to a false inclusion of a specific individual.

DYNAMIC DISPLAYS

As with printed images, dynamic images are composed of pixels. In most cases, each pixel is composed of three smaller patches, known as subpixels. Each subpixel emits light of one of the primary colors. In some cases the subpixel passes the light, and in other devices, it actually produces the light. In the case of printed materials, the primaries that are used are the subtractive primaries: cyan, magenta, and yellow. In the case of dynamic displays, they utilize the additive primaries: red, green, and blue. The subtractive primaries are embodied in filters or dyes, and they are viewed under white light. Each dye subtracts one of the additive primaries. What is left are various amounts of light from each of the additive primary colors. Visually, the person viewing the image adds the contributions of each of the lights. In the case of the additive primaries, the system presents the additive primaries directly. Again the viewer integrates the resulting combinations of lights. Remember that the additive primaries are lights, and the subtractive primaries are dyes or filters. The relationship among these is that each of the subtractive primaries blocks one of the additive primaries. Cyan blocks red, magenta blocks green, and yellow blocks blue.

As with printed images, the higher the number of pixels, the higher the resolution of the image. Close examination of any dynamic display will show this. For example, look at a gray patch on your computer screen through a magnifying glass and you will see very small groupings of red, green and blue spots. Each such grouping is a pixel. On some monitors, the groupings are triangular arrangements of colored circles (separated by the black of the shadow mask). On others they may be three vertical lines placed very close to each other.

COMPUTER-DRIVEN DISPLAYS

There are five technologies that are commonly used to produce computer-driven displays. The oldest of these is based on the Cathode Ray Tube (CRT), but sales of these devices are decreasing as three other technologies have come to the fore. Notable examples are Digital Light Processing (DLP) technology; Liquid Crystal Display (LCD) technology; and Plasma Display Panels (PDP). More recently, Organic Light-Emitting Diode (OLED) technology is starting to find a place. Performance of the different devices is based on several factors, including:

- **Resolution**. The ability to display fine detail as a function of the screen size.
- **Brightness**. How bright the whites are, usually measured in lumens.

- **Contrast ratio**. How dark the blacks are compared to the brightness of the whites. This usually is shown as a ratio. This is not the same as the image contrast, which depends on the content of the image. This contrast ratio is purely a function of the device under test.
- **Size or desktop footprint**. How big is the physical device as a function of the actual size of the display screen?
- **Flat screen capability**. Can the technology be configured into a relatively flat screen?
- **Ability to project**. Some image-forming technologies can be used in projectors and others are not well suited to this application.
- **Screen size**. How big can the screen be?
- **Responsiveness to fast-changing signals**. Can the display respond well to rapidly changing input signals?
- **Power consumption**. The watts consumed per unit of area of screen in a given application (direct view or projection).
- **Viewing angle**. Does the user have to look straight on to the display or can he or she be viewing at a substantially oblique angle?
- **Artifacting**. Does the device introduce visual features that are not proper parts of the image being displayed?
- **Lifetime**. How long can the device be expected to last?
- **Cost**. Given values for all of these, how much do the devices cost?

Cathode Ray Tube (CRT)

As the name implies, CRTs operate by sending rays of electrons from a cathode inside a tube down to a screen of phosphors on the inside of the front face. The phosphors convert the energy of the electrons into visible light. The device has integral coils to focus the beam of electrons, and either electromagnetic coils or electrostatic plates that are used to steer the electrons to a series of selected locations on the phosphor screen.

The electrons come from a device called an electron gun (refer to Figure 8.4). This is a device that, when heated, emits a stream of electrons. It has a control grid at its muzzle, which controls the rate of electron beam production. That is, the grid can fully open the gates and let out a large number of electrons per second, which gives a bright spot on the phosphor screen; or it can shut it down, giving virtually no light emission. It is capable of a full range of intensities between these extremes.

The beam can be deflected independently in the vertical and horizontal directions very rapidly, and as this happens, it moves the spot of light to proper locations on the screen. In many devices, the horizontal deflection signal moves the beam steadily across the screen from one side to the other, and then drops back very rapidly to the initial setting. So the beam

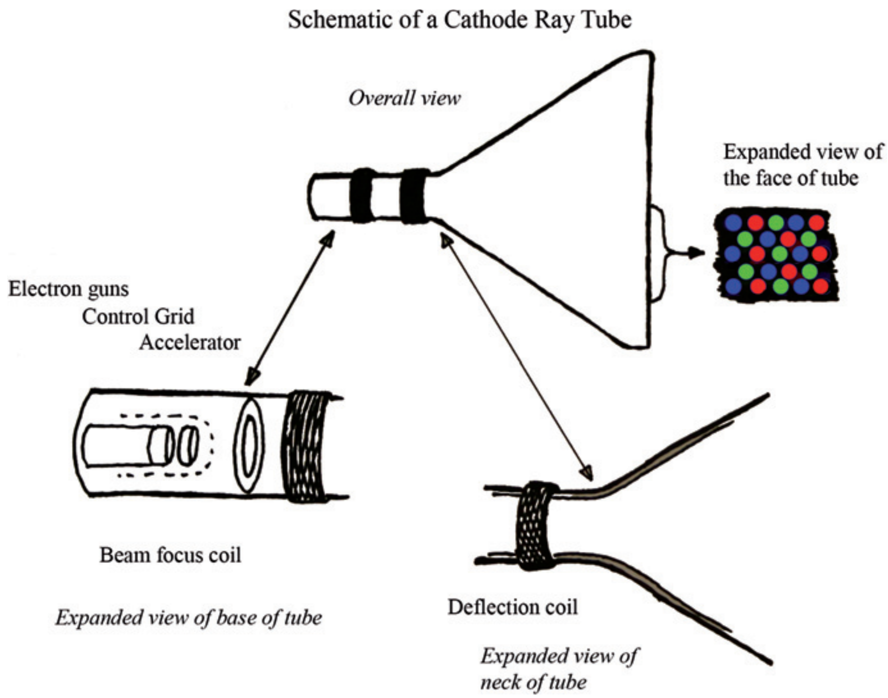


FIGURE 8.4 *The Cathode Ray Tube. The overall schematic is shown at the top. Key parts of it are shown in blow-ups below and to the right. At the base of the tube is the electron gun, which liberates electrons, forms them into a beam, and sends them down the tube. At the neck of the tube this drawing shows a coil that causes the electron beam to move across the face of the tube. Some CRTs use electrostatic deflection plates instead of a coil. The blow up of the face of the tube indicates the triads of spots of phosphor and the black shadow mask.*

seems to move from left to right, and then disappear, only to reappear on the left again and repeat the sweep. The wave form of this signal is called a sawtooth. In waveform display devices, such as oscilloscopes, there is no preprogrammed vertical deflection. This is driven by the input signal. In computer monitors and televisions, there is a preprogrammed vertical deflection signal. Traditionally the spot starts in the upper left-hand corner of the display screen and moves across to the upper right-hand corner. Then it disappears and reappears on the left again, but just below its first sweep point. Each sweep from left to right is called a row, and each transition from right to left is called a flyback.

In color displays, there are three electron guns, and three beams move down the tube to the screen. The screen is coated with spots of red, green, and blue phosphor material interspersed by a sieve, or shadow mask, that separates the individual spots. This ensures that the red phosphor is activated only by the red intended beam, and so on. The spots appear in groups of three: red, green, and blue (one of each). The most common arrangements are small, closely spaced vertical bars, or circular spots in a triangle. Each triad of spots constitutes a pixel, and each of the elements within the pixel is a subpixel. As in all of digital imaging, the spacing of the pixels per unit of length, or per screen dimension, indicates the device resolution. Since CRTs have a fixed pattern, the maximum resolution is this fixed value. The device can be programmed to work at integer fractions of this value—½, ⅓, ¼, and so on—but not at higher values. CRTs can be made with rather high resolutions, but in large formats they become prohibitively expensive, cumbersome in size, and consume huge amounts of power. CRTs produce their light essentially at the screen face, so they can accommodate a wide viewing angle. CRTs are not as suitable for projection as some alternatives, like DLP and LCD, since these have a light source separate from the screen. Conversely, when used as a direct-view device, they—like Plasma and OLED devices—are relatively bright.

The phosphors can "burn in" over time if certain parts of the screen are used repeatedly to produce high levels of output. This causes a ghost image and is the reason behind the "screen saver." Also, the shadow mask can become magnetized. This will cause noticeable and fixed patterns in the image. Most of the newer CRTs have a built-in demagnetizer to reduce any magnetization every time the device is started up. The phosphors have a relatively short persistence. That is, they do not continue to produce significant amounts of light after a few milliseconds. The persistence time is set to coincide with the time required to produce an entire frame. There is a very low level of afterglow that can be seen when a TV set is turned off. The result is that CRTs have a relatively good contrast ratio.

Plasma Display Panel (PDP)

The design of PDPs is similar to CRTs in that a phosphor on the inside of the front face creates the light. As such they can achieve good brightness and can be viewed over a wide range of angles. The difference is that instead of using electrons (as in CRTs), PDPs create inert gas ions and, in the process, produce UV photons. The UV causes the phosphor to glow (much as in a fluorescent lamp). Also, instead of accelerating the ions down the length of a glass tube, the excitation is created in a small cup just behind the phosphor screen. Referring to Figure 8.5, each cup (rectangular in the drawing) is a subpixel, and groups of subpixels comprise pixels. There are vertical and horizontal conductors that are activated individually to select a specific location—that is, a specific subpixel. The amount of current allowed to flow through each selection controls the brightness of the light produced.

One advantage in the design of PDPs is that when a given cell is not turned on, it goes quite dark. The result is that they have good contrast ratios, and since the current through each cell controls the light level, very good tone reproduction can be achieved. They are good for photos and movies.

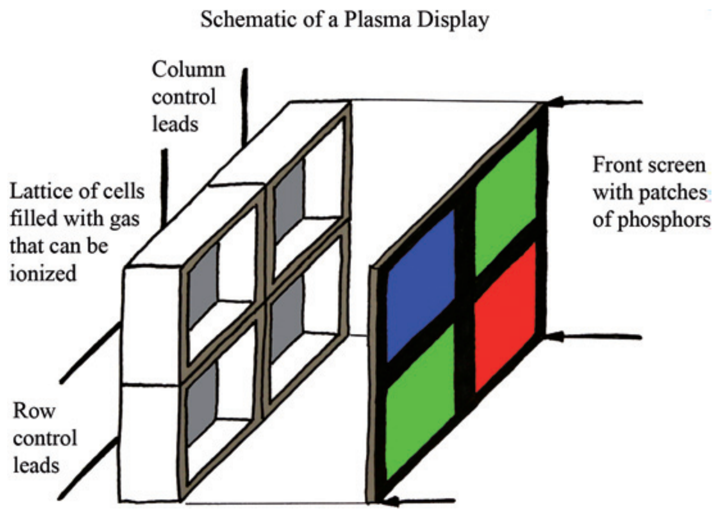


FIGURE 8.5 *The Plasma Display. Four of the millions of small cups are shown. Each is independently driven and each holds a small amount of gas. When activated, the gas is converted to plasma, which excites the phosphor coated on the inside of the front screen.*

PDP technology is very good for large screens. They can be as thin as 4 inches in depth. They require power levels comparable to CRTs. Also like CRTs, prolonged, high-brightness applications from certain pixels will cause the phosphors to lose some of their brightness, resulting in burn-in effects.

Liquid Crystal Display (LCD)

LCDs have enjoyed a steady increase in quality and performance over the past few years and are beginning to usurp the once-dominant position of PDPs in the large screen segment. They are also popular in smaller formats and this is usurping the position of CRTs. LCDs do not use phosphors. Instead the light is provided by a general source that illuminates all of the pixels. Each pixel has the ability to either pass or reflect this light or not, depending on individual control signals. In reflective LCDs, the ambient light is reflected back to the viewer; and in transmissive LCDs, there is a light source behind the layer of pixels and this light is either passed or not. LCD watches use reflective LCDs and televisions use transmissive LCDs.

The key process involves crystals that are birefringent—that is, they rotate the angle of polarization of light that passes through them. Moreover, when these are exposed to an electric field, the angle of rotation is decreased. To make an LCD, several small areas (or subpixels) are created, each of which contains these crystals in a suitable medium (refer to Figure 8.6). The crystal medium is sandwiched between two layers. In reflective LCDs, one of the layers is transparent and the other is reflective. In transmissive LCDs, both

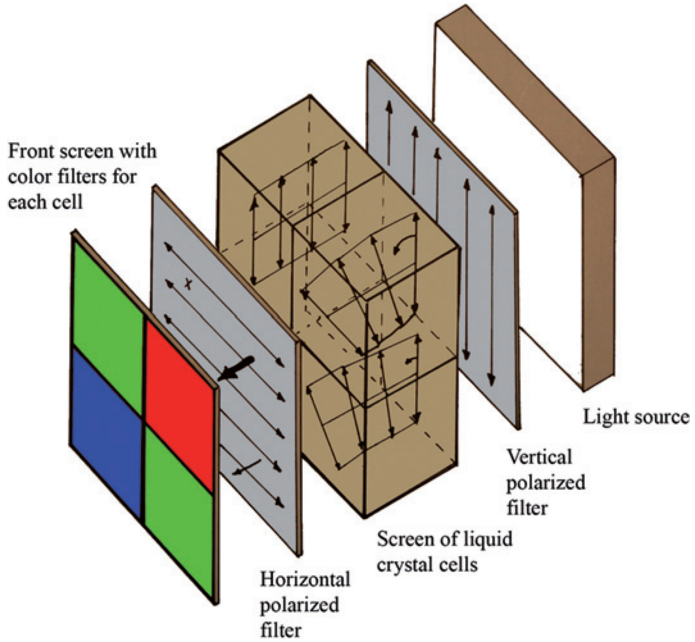


FIGURE 8.6 *The Liquid Crystal Display. Light is provided by a uniform source at the back of the display. It passes through a polarizer that passes primarily vertically polarized photons. These enter a sandwich of two sheets of individual electrodes with a liquid crystal layer between them. When a particular pair of electrodes is activated, the liquid crystal material rotates the angle of polarization. This light is now able to pass through the horizontal polarizer and pass through a color filter on the front screen. With no activation, the light is not rotated and the element is dark. With partial rotation, some of the light comes through.*

layers are transparent. In addition to the two layers just noted, there are two polarizing filters. One is on the input side and the other is on the opposite side. One is set to polarize the light horizontally and the other, vertically. Consider the transmissive mode. Light comes in one side and is polarized vertically. It goes through liquid crystal material and the angle of polarization is rotated 90 degrees. It now passes through the second polarizer, which is set to horizontal polarization. If an electric field is applied to a given pixel, then the rotation by the liquid crystal material is not imparted and the light does not emerge polarized vertically. Thus this light will not pass through the second filter. The electric fields are applied by means of transparent electrodes that are coated on the two layers that contain the liquid crystal material. Mid levels of brightness are achieved by using lower electric fields, which do not change the polarization as much as the full field.

In the transmissive case, a light source is provided behind the light modulation sandwich. Typically this is a fluorescent source, and if the display is a color display, each subpixel is coated with a colored filter—one of the additive primaries. Since all the light has to go through at least one polarizing screen, even in the *fully on* mode, the light from the source is reduced by at least a factor of two. To make the screen brighter, as in projectors, the light must be greatly increased. When a pixel is set to *off* by means of the application of the electric field, there is still some light that leaks through. The result is that LCDs have lower contrast ratios than either PDPs or CRTs. They also have a smaller color gamut.

LCDs can have a noticeable delay in reaction to the application of the electric fields, and sometimes there are visual indications of this. LCDs also perform best when in their native resolutions. If the screen has 1080 pixels across the screen, it is best to operate it at that level. Because of the light paths involved, LCDs tend to have more restrictive viewing angles. Since fluorescent lights are rather efficient, and since current flows through the pixels only while they are being switched, LCDs use much less power than CRT or PDP devices. This makes them attractive for battery-operated devices, such as viewers on the backs of digital cameras and cell phones.

Digital Light Processors (DLP)

DLPs are composed of an array of tiny mirrors hinged on a small, flat platform called a Digital Micromirror Device, or DMD. Each mirror is a subpixel and is driven by an electromechanical actuator. Light is projected onto the mirror and it either reflects that light to the viewing screen or onto a heat sink called a *light dump* to be absorbed. The mirrors are extremely small and can be rotated very quickly, thus the mirror can be made to flutter on and off many times during a single TV frame. The viewer can see only a single frame, and so the apparent brightness of any subpixel is controlled by the amount of time that the given subpixel shines its light on the screen instead of on the light dump.

DLP technology was developed by Texas Instruments, Inc., and the term still refers to their devices. They have excellent contrast ratios, since when the mirror is reflecting light to the screen, that light is essentially as bright as the original source (as filtered through a red, green, blue, or other filter). When the mirror shines on the light dump, no light arrives at the given spot on the screen.

DLPs come in three- and one-DMD configurations. In the three-DMD units, three light beams are created and each carries one of the three additive primary colors. These shine on their own respective DMDs. The reflected light from all three is then combined onto a single screen. In a one-DMD unit, the light source shines through a filter drum or wheel on its way to the DMD. Thus the three color fields alternate. This reduces the amount of time for the mirrors to flicker on and off to modulate and create the brightness levels. These devices are prone to showing the "rainbow effect," which is a bit of saturated color streaking into the image. The three-DMD units do not show this effect.

Since DLPs require that a light source shine onto the DMD and then onto the screen, the device must have room for an optical path. The result is that DLPs are not nearly as flat as LCDs or PDPs but they are more compact than CRTs. DLPs with three-DMDs are excellent displays, but they are a bit more costly than LCDs these days and their market share is being reduced. They make excellent projector units and are well represented in that market. DLP technology with three-DMDs is being used in movie theaters where they give excellent tone scales, vivid colors, excellent contrast ratios, and are capable of many more colors than most technologies.

Organic Light-Emitting Diode (OLED) Displays

OLEDs are assembled in a fashion similar to that of LCDs in that there is a layer of active material sandwiched between two structural layers. Unlike LCDs however, there is no need to shine a light through the full layer, so the backing will always be reflective and the front layer will be clear. There is a cathode (a back layer that can supply electrons), a conductive layer, an emissive layer, and an anode (a positive electrode layer). The light comes from the emissive layer and the material selected for each subpixel is chosen to give one of the three additive primary colors.

This technology was developed by the Eastman Kodak Company and was developed for use as a display on the backs of digital cameras. These devices have very high light output per watt of power consumed, so they are ideal for battery-operated devices, such as digital cameras or cell phones. They are also easy to manufacture since the various layers can be laid down onto substrates much the way a printing press makes the pages of a book. Unfortunately they do not yet have good enough lifetimes for many applications. If the longevity can be improved, they should enjoy success for low-power, high-brightness, and low-cost displays. Sony Corporation has just announced a line of television sets that use OLED technology.

Comparisons

Although the real comparison should be for each technology relative to an application, Table 8.1 relates the relative performance characteristics of the technologies described.

Although these technologies are undergoing development and change, they can be roughly rated per the chart. Better performance is indicated by + signs, *more* being better, and a − sign indicates where performance is not high.

For courtroom use, CRT or small LCD devices would serve individual juror viewers well. For a general view, such as a wall-mounted screen, a PDP device would serve as a good choice. If projection onto a general view screen

Table 8.1 *Performance Comparison*

Factor	CRT	PDP	LCD (Relative Strength)	DLP	OLED
Resolution	++	+	+	++	++
Brightness	+++	+	+	+++	++
Contrast Ratio	++	++	+	+++	++
Footprint Size	+	+++	+++	++	+++
Flat Screen	–	+++	+++	+	+++
Projection	+	+	++	+++	+
Screen Size	+	+++	+++	+++	++
Responsiveness	+++	+++	+	+++	++
Power Usage	+	+	++	++	+++
Viewing Angle	+++	+++	+	+	++
Artifacting	+++	+++	++	++	++
Life	+++	++	++	+++	–
Cost	+	++	+++	+	+

is desired, DLP technology would be a good selection. For use on portable devices such as cameras, LCD and OLED (assuming improvements in life) should be chosen because of low power consumption.

When it comes to making hard-copy prints, the new, high-performance inkjet technology is most likely to be the best choice. There are printers that can make relatively large (poster-sized) images suitable for general display. For individuals, there are many high-quality devices available. The newer printers have a relatively wide range of colors, relatively high-density capability, and high resolution. There are devices that write with lasers onto traditional photographic paper, and these can give excellent results as well. Dye sublimation printers will give very good images, but there are print size limitations. Low-end inkjet and laser printers will give less than optimal images.

Key Photographic Techniques

There are many special techniques that have been developed by photographers over the years and it is good to be aware of the more common of these. Almost all were developed based on traditional film photography, but they easily are adapted to digital camera photography. These techniques are discussed in this chapter.

PAINTING WITH LIGHT

This technique is used to photograph large scenes in very dim light, where it is not feasible to illuminate the entire scene with a single light or with a collection of many lights. This technique requires two people: one to operate the camera and one to operate the flash unit, which is removed from the camera. The exposure is computed using the flash output guide number and the distance from the flash to the subject. Modern flash units might have the ability to indicate an f/stop for given distances, and this makes the process easier. The camera is set to a fixed ISO and the indicated f/stop is set manually. The camera is placed on a tripod and the shutter speed is set to either the "time" or "bulb" setting to keep the shutter open.

One person stands at the camera with a black card and covers the lens between individual flash exposures to ensure that minimal extraneous light enters the camera. The card is removed during the actual exposure. The other person moves around the scene, strategically lighting each segment individually with a hand-held flash attachment. Care must be taken to keep the distance from the flash unit to the part of the scene being illuminated consistent for each successive flash exposure. Also this person must stay out of the line-of-sight of the camera for the particular area being exposed. This technique requires some experience and some trial-and-error, but it is very good for lighting large crime or accident scenes at night. The process is a bit more difficult with digital cameras because they have less dynamic range and because dark areas tend to be noisy—especially for long exposures.

MACRO PHOTOGRAPHY

This technique is used to photograph extreme close-up images of small objects when images that are the same size as the object or larger are required. To accomplish this, special-purpose macro lenses (some manufacturers call them micro) have a long barrel for the close focusing that is required. Some macro lenses are capable of magnifications up to 5:1; however, it is more common for a photographer to use a "standard" (1:1) macro lens. There are less expensive means of doing macro photography, such as extension tubes, a bellows, or close-up lenses. But most photographers prefer the macro lens because of its sharpness and ease of use.

Some other special considerations apply when doing macro photography. The depth of field is very limited when working at close distances and must be controlled by the f/stop. This means that focusing is very critical and must be done on the most important part of the subject with the camera mounted on a tripod. To increase the depth of field, high f/stops are used, which implies longer exposure times. To ensure no camera movement during the exposure, a cable release should be used. Sufficiently and evenly lighting the subject can also be a large problem and difficult to overcome. Some cameras can focus on subjects so closely that they almost touch the front of the lens. It is impossible to correctly light a subject that close. Avoiding this problem may require the use of longer focal length macro lenses, such as in the range of a 100 or 200mm lenses. This will increase the camera-to-subject distance allowing more room to properly light the subject. Most macro lenses have been designed for use on 35mm cameras, and putting them on some digital cameras can result in a misalignment between the area imaged by the lens and the image captured by the camera.

ALTERNATIVE LIGHT SOURCE (ALS)

The ALS is used for bringing out images of materials that are either not visible to the eye, or are mixed into a distracting background. The process relies on fluorescence. Typical applications include latent fingerprints, blood, fibers, bodily fluids, and other trace materials that are difficult to see under normal light conditions and ones that either naturally fluoresce or can be made to do so. Most ALS systems are adjustable, making them capable of using several wavelength bands of light from approximately 300 to 700 nanometers. Exposures are made at short wavelengths to excite the fluorescence response; then the image is recorded at the longer wavelengths due to the fluorescence emission. The light from the fluorescence will be at a longer wavelength than the exposing light because it is at a lower energy level. A filter is placed over the lens of the camera to pass the fluorescence light and block the light used in the exposure. In this way, a bright image is seen on a dark background. The nature of the substance that is being photographed will dictate the color of light needed to excite fluorescence and the color of the emitted light. Earlier digital cameras did not have sufficient sensitivity to make ALS images, but the newer cameras do.

THREE-POINT LIGHTING

This is the most common lighting technique used in both still and video photography. The photographer can light the subject, while also controlling the shadows created by direct lighting. The three lights include a key light, fill light, and back light. The key light (sometimes referred to as the main light) illuminates the subject and determines the shot's overall lighting design. It is usually placed in front and to one side of the subject. The fill light shines on the subject also, but from the opposite side of the key light, and is usually less bright than the key light. It should balance the key light, reducing or eliminating the shadows cast by the subject. The back light (sometimes referred to as a hair light) shines on the subject from behind, and in some cases, to the side. This light tends to separate the subject from the background. A fourth light called a background light is sometimes added, making this a four-point lighting setup. The background light is placed behind the subject and is used to light background elements of the photograph. The background light is used frequently to eliminate shadows cast by objects in the foreground. Since digital cameras have less dynamic range than negative film systems, this approach to lighting a subject can be very advantageous.

DIRECT FLASH

Most people are familiar with direct flash—the flash is mounted on the camera and is fired directly at the subject. This is found on all cameras that have built-in flash units, from simple point-and-shoot cameras to consumer-grade SLRs. Direct flash photographs are easy to recognize. The photographs will have harsh shadows directly behind the subject, the contrast of the scene will be high, and the photos will have very little appearance of depth. If the built-in flash is mounted close to the axis of the lens, red-eye will also be a problem when photographing people in subdued light.

This is the least desirable technique when photographing people. It can also be a major problem is there is a highly reflective material behind the subject, such as a window, a mirror, or a sheet of bare metal. In these situations, the light from the flash will reflect directly off the background and overpower the subject—especially if automatic exposure is used. Most flash units are dedicated to the camera on which they are mounted and exposure will be computed correctly by the camera if you are using the program mode and are within the range of the flash. Digital cameras, with their reduced dynamic range and tendency to bleach out bright detail, are particularly sensitive to this type of lighting.

DIFFUSED FLASH

This method is used to soften the harsh lighting associated with the direct flash technique. A diffuser, such as a translucent sheet of plastic or cloth, is placed over the flash attachment. Some flash units have built-in diffusers and can be pulled down over the head of the flash. Some light will be lost, but the camera will compensate for the loss of light when computing exposure if you are within the range of the flash. This is particularly valuable with digital cameras and subjects that are close to the camera.

BOUNCE FLASH

This technique can be used to soften and even eliminate harsh shadows caused by direct flash while retaining the convenience of on-camera flash. Bounce flash units have the ability to angle the flash head upward in order to reflect light off a ceiling or a reflector connected to the flash unit. One thing to keep in mind when using bounce flash is the angle of the light. Light reflects at the same angle at which it is incident, so the photographer should visualize where the light will hit the ceiling and follow the same angle off the ceiling to be sure that it illuminated the subject. It doesn't have to be exact, but care should be taken to assure the light doesn't drop down too far in front of or behind the subject.

Some light will be lost in the process of being reflected, typically about two stops. When using a dedicated flash on a camera with TTL metering, the camera will make the correction for you. The ceiling is a common choice for the reflector because it is usually white and not glossy. Also, since normally we see an object illuminated with light coming from above, it makes the subjects appear normal. If the ceiling is not white, or is more like a mirror than a flat painted surface, this approach will be less satisfying.

FILL FLASH

Fill flash typically is used to add illumination to the foreground to help balance the bright light levels of the subject and the shadow areas so both can be captured in the same frame. This can be used to brighten the subject when it is lit from behind (backlight), and often is used outdoors. Some situations where this technique is useful are:

- A harsh midday sun creating heavy shadows on the subject
- Snow or a light sandy beach behind the subject

- Light coming from a window into an otherwise dark room, and from behind the subject

Most of the later-model cameras with dedicated flashes and TTL metering have a setting for fill flash, which will automatically compute the exposure. Beware of over-powering the main light because the photograph will not look real. This approach is another good way to compensate for the lower dynamic range afforded by digital cameras.

OFF-CAMERA FLASH

This process is used for lighting an object at an oblique angle. It is the best technique to show surface textures. In this technique, the flash attachment is placed somewhere off to the side of the camera. A number of methods can be used, but most commonly it can be achieved with the normal flash attachment, using an extension cable and possibly an adapter. Alternatively, it can be done with an on-camera flash and an off-camera slave flash (which is triggered by the on-camera device). With the extension cord method, there is a physical connection between the camera and the flash unit's hot-shoe, which limits the distance of the flash. But if the flash is TTL dedicated, exposure can usually be controlled by the camera.

Another method is to use a "slave flash." These units have built-in sensors that are triggered when they detects another flash discharging. But there is no camera-to-slave communication, so usually it is not possible to use TTL function on the camera. Most slaves must be manually adjusted and charged. Still another method is the use of a wireless TTL flash. This refers to a sophisticated infrared-light-based, or radio-wave, camera-flash communication that provides full flash functionality "at a distance." Most wireless units will support both manual and automatic flash, but TTL mode is the most desired. Since most of the new cameras being designed today are digital, the technology to support the wireless, off-camera flash will develop more rapidly in the next few years.

RING LIGHT

In close-up work, this reduces sensitivity to surface roughness and provides very even lighting on the object. A ring light is a flash head mounted on the front of the lens that is shaped like a doughnut. The shot is made through the doughnut hole. The flash tube creates a continuous circle around the lens, thereby providing even illumination from all angles around the camera's optical axis. Actually there are two types of ring lights: the flash type that was just mentioned and another that contains LED lights. They both appear similar in construction but have different output capabilities. The flash ring light is much more powerful in most cases.

Exposure is not a problem; if the camera has TTL metering, it will automatically compute the exposure for the LED light. The flash ring light is dedicated and the TTL metering will set the exposure. Use of a ring light is indicated to achieve just the opposite effect of that of off-camera flash. Off-camera flash is used to enhance surface texture and ring light is used to subdue surface texture. The new technology involves the use of LEDs, and this is expected to develop rapidly in the next few years. It is already finding wide use in police car emergency light bars.

45°–90° ILLUMINATION

This type of lighting commonly is used to minimize front surface reflections from the object. It refers to how the lights are positioned in relation to the object being photographed. It typically is used on copy stands, but also applies when one is using off-camera flash. A normal setup would have the camera at 90° to the object and the axis of the lights set to 45° relative to the object's main surface. The angle will always be determined by the surface of the object. In some cases the angle may be less than 45°, depending on the surface being photographed. This is the standard for flatbed scanners and measuring instruments, such as reflection densitometers, as well. There is virtually no difference in using this technique with either film or digital cameras.

PERSPECTIVE GRID

This is used to determine three-dimensional information from normal photographs. It typically is used at crime scenes and accident scenes. A flat square of known size is placed in the scene, usually on the floor or ground in front of the camera, and is used to extrapolate to vanishing points as well as to establish a basis for making measurements in the horizontal plane from the image. To get information regarding the vertical, we would need to place a similar square on a vertical wall. There are often elements in the scene that can serve the same purpose as the squares, assuming their sizes and orientations are well known. New devices based on infrared scanning technology are now available to accomplish this function, and though they are currently quite expensive, costs will probably come down in the future.

INFRARED PHOTOGRAPHY

This type of photography allows you to see by a light that has a wavelength of 700 nanometers or more. This light is not visible to humans and can be from either a hot light source or heat from within the subject itself. It is sometimes referred to as *thermal imaging*. Infrared photography makes it possible to "see" a subject with or without any visible illumination—you can see in the dark. It also lets you see variations in temperature, thereby determining the temperature at various positions on the subject.

Photographic film does not have sensitivity native to this portion of the spectrum and special sensitizers need to be incorporated to make films responsive to infrared light. Digital cameras have the reverse situation. Silicon has high native sensitivity to near infrared light and normally an infrared blocking filter has to be included to ensure that what the camera is seeing as red, green, or blue does not include infrared response as well. So for film photography, you need to use special films with infrared sensitivity; and in digital photography, you must have a camera that allows removal of the infrared blocking filter.

When using film, an infrared-passing filter is used to let infrared (IR) light pass through to the camera and blocks the visible light portion of the spectrum. A typical outdoor scene will probably have dark skies and the clouds will really stand out. A person's skin tends to look milky, although his or her eyes often look black. Special considerations must be made to focus the camera. Many of the manual focus 35mm SLRs have a red dot, line, or diamond, called the *infrared index mark*. This is used to achieve proper infrared focus. Some of the new autofocus lenses no longer have the mark. Using the infrared filter, which is opaque to visible light, the SLR becomes useless for both framing and focusing. Place the camera on a tripod and compose the photograph without the filter, then install the filter. If you do not focus the camera meticulously to the infrared index mark, a sharp image might be obtained by using a small aperture (f/22) and a slow shutter speed.

In the case of digital photography, a special camera is required. Focusing and framing are helped by the viewer on the back of the camera and by the auto-focus capabilities of the camera. The red, green, and blue filters over the pixels are transparent in infrared, and so this is not an issue. The important feature is the ability to avoid any infrared filter built into the camera to facilitate normal, visible-light, color photography. There are new camera models that are designed specifically to capture images at wavelengths ranging from the near UV (350 nm) all the way to the near IR (1000 nm).

ZONE SYSTEM

This system is used for determining optimal film exposure and development in order to maximize the impact of a scene's range of light levels. It was developed relatively early in the history of photography by Ansel Adams, who used it to achieve outstanding photographs. It provides photographers with a method of precisely defining the relationship between the way they visualize the photographic subject and the final results. It all starts with exposure metering. An averaging meter, which is most common, cannot distinguish between a subject of uniform luminance and one that consists of light and dark elements. Using the zone system, measurements are made of individual scene elements, and exposure is adjusted based on the photographer's knowledge of what is being metered.

A photographer knows the difference between white objects and black objects—the light meter does not. The brightness levels of the key parts of the scene are categorized into zones, and then the exposure is set to locate these key parts on the usable exposure range of the film or sensor. For every film, developer, and paper, there is a "normal" development time that will allow a properly exposed negative to give an accurate print. In general, optimal negative development will be different for every type and grade of paper. All digital cameras have built-in metering and usually have means to allow the photographer to select the portion of the image that has to be rendered with good tonal detail. They also have limited dynamic range, so trying to implement a zone system process will be difficult. Digital photography facilitates extensive and simple postprocessing and, to a large degree, the impact sought by the zone system can be achieved by other means.

COMPONENT PANORAMA

This is a process for making a photographic representation of a very wide scene. Several photos are taken, each at a different angle, pivoting around a single point where the camera is usually mounted on a tripod. Care must be taken to ensure that the images overlap enough to be properly joined, or "stitched" together, to create a single image. The resulting images are panoramic because they offer an unobstructed or complete view of the full area. The image usually takes the form of a wide strip. Panoramas photographed with a digital camera usually are stitched together using software on a computer. Most of the newer digital cameras come with stitching software. When shooting a panorama photo and pivoting the camera, some points will be relatively close to the camera and others will be farther away. The close objects will be large in comparison to the ones that are farther away. The result is a variation on a theme of fish-eye distortion.

Another version of panorama photography is to move the camera and take a series of photos along the way. For example, with a tire track, you could put the camera on a tripod, set it perpendicular to the ground, and take a picture. Then move the camera along the track, check to see that the camera-to-ground distance is consistent, and take a second picture. This process can then be repeated several times along the track. The resulting series of photos can then be stitched together to form a long, high-resolution rendering of the tire track.

With film photography, you can take a series of photos, scan them, and then use the computer software to stitch them together. With digital photography, the images are ready to go right out of the camera.

PUSH PROCESSING

This is a film photography approach and refers to a development technique that increases the speed of the film being processed. Using the push processing technique lets you increase the film speed above the manufacturer's recommendation. This will allow making photographs under lighting conditions that ordinarily would be too low for good exposures. Basically the film is underexposed and overprocessed. Using this technique comes at a cost, though. Usually there is an increase in grain, an increase in contrast, and a decrease in quality, at the very least.

Another technique is pull processing, which is just the opposite. This decreases the speed of the film and is achieved by developing the film for a shorter time. Digital cameras have an ISO adjustment that is intended to make the camera more or less sensitive, depending on where the dial is set. As with films, however, selecting very high sensitivities can result in a degradation of image quality. In addition, it is easy to adjust brightness and contrast in such a way as to simulate an increase in sensitivity. This will take a toll on image quality as well.

BRACKETING

This is a technique used to help ensure a correct exposure when dealing with conditions that are difficult to measure with a light meter. It is done by estimating a suspected proper exposure and then taking a series of photos at differing exposure levels—some below the suspected exposure and some above. The idea is that at least one of these will be right. It was very common with film because one couldn't see the photograph at the scene. But it is not really necessary with digital cameras, because one can instantly view the pictures at the scene. With the better digital cameras, one can even see a histogram of the image just taken and make expert corrections before taking a second shot. Most of the late-model film and digital cameras have a bracketing mode built in to the camera.

At the very least, three shots are taken of each scene—one at the suspected camera exposure, one over-exposed, and one under-exposed. This can be set up in the camera to determine how much the exposure should change between pictures and how many pictures should be taken of each scene. If the camera does not have it as a built-in mode, manual adjustment of exposure will be necessary.

LENS TRANSLATION

This is a process for obtaining photos with the 90° perspective when there is an obstacle that prevents holding the camera directly above or in front of an object. Using a flexible bellows device, the camera lens is moved laterally off to the side with its axis still normal to the object's principle surface. This requires a bellows extension with lateral adjustment. Corrections must be made manually to any exposure settings on the lens because it normally cannot be controlled by the camera's exposure system through the bellows attachment. It also requires a lens with low barrel or pincushion distortion and a wide field of view capability. When using this process, there is no significant difference between film and digital photography.

POLARIZED LIGHT

Normal light is composed of waves that vibrate at all angles around the axis of propagation. When the angle of vibration is restricted, the light is said to be *polarized*. There is single-angle polarization and circular polarization. When light reflects off certain materials, such as metals and water, it comes off polarized. There are filters that can polarize a beam of light—that is, the light coming through the filter is polarized. By rotating the filter, you can select the plane (angle) of polarization. This process is used to reduce certain front surface glare components when taking photographs.

A special polarizing filter is placed over the camera lens and rotated until front surface glare is minimized. For example, consider taking a photo of someone in a pool where there is glare coming off the water. The water will tend to reflect light polarized at a particular angle. The filter on the camera is rotated to the point where it is not allowing in light at that particular angle of polarization. The result is a noticeable reduction in the glare off the water.

There are two types of polarizing filters: a linear polarizer (which passes light vibrating in only one direction) and a circular polarizer (sometimes called a CPL filter). The metering and auto-focus sensors in certain SLR cameras will not work properly with linear polarizers because of the mirror and the beam-splitters used to split off the light for focusing and metering. Circular polarizers will work with all types of cameras.

CONTRAST FILTERING

This technique is commonly used in black-and-white photography as a means for increasing or decreasing contrast among elements of the object that are of different colors. Colored contrast filters are used over the camera lens to select certain elements compared to others. The result is an increase in contrast according to the elements' basic colors. Contrast in items of the same color of the filter will be decreased, and that of opposite color will be increased. For example, the use of a light yellow filter when taking a picture of white clouds against a blue sky will result in increased contrast in the sky and decreased contrast in the clouds. Using the correct color filter with black-and-white film can sometimes completely eliminate a color from the image. Since digital cameras intrinsically take color photos, this technique is of limited value.

STEREO PHOTOGRAPHY

This is a means for showing depth in photography. It requires the use of either two cameras, or a single camera with two lenses. The lenses are slightly separated (as are a person's two eyes). The two separate images are printed side by side. If viewing them without a stereoscopic viewer, you are required to force your eyes either to cross, or to diverge, so that the two images appear to be three images. Then as each eye sees a different image, the effect of depth is achieved in the central image of the three.

Obviously it is much better if a stereoscope is available. Usually the stereoscope has magnifying lenses that change the focus point of the image from its short distance to a virtual distance at infinity. It offers a wider field of view and a partition between the images, avoiding a potential distraction to the user. This can be used with either film or digital technology. It is also possible to project the two separate images through separate projectors and place a polarizing filter over each projector lens. The two are set to different angles. The viewer wears a pair of glasses with polarizing filters consistent with those used on the projectors. Then, even though the two images are superimposed on the screen, the two eyes see the separate images separately and the brain puts the combination together as a stereo image.

THREE-DIMENSIONAL SCANNING

This technique provides a means for recording a three-dimensional photograph of an object with a single device. A special scanner is set up on a tripod, which takes a normal photograph while also sweeping the surrounding environment with a pulsed laser beam at an angle to the optical axis of the camera. The laser rays produce reflections when they encounter solid objects and the image produced is an array of dots called a *dot cloud*. The device knows the angle of the laser beam at each point, and if the object were a flat plane, all the dots would be exactly where they would be predicted to be from the angle of incidence. If the object has parts that are not in the flat plane, the dots will be displaced to the left or right of where they would have been on the plane.

Using simple trigonometry, you can calculate how far in front or in back of the base plane the object's surface was. The instrument determines the XYZ coordinates of each reflection point and can create a three-dimensional mathematical model of the object. Using this technique, you can scan a solid object from several different directions and later merge the computed partial surfaces into a single, continuous representation of the three-dimensional object. Spatial interpolation of the reflection points yields a three-dimensional surface model of the object. The device then attaches the normal photo to render the object's surface markings, called *texture*, to compute a three-dimensional model of the object, complete with surface graphics. This technique will work with film photography in theory, but only digital camera versions are commercially available.

PHOTOMICROGRAPHY

This refers to the capturing of images at very high magnification. A special adapter is placed on a microscope to allow a camera to receive the image. The quality of a photomicrograph, digital or film, is dependent upon the quality of the microscope. The microscope is configured using Köhler illumination, and the field and condenser diaphragms should be adjusted correctly. When properly adjusted, the microscope will yield images that have even illumination over the entire field of view and will display the best compromise of contrast and resolution. Most modern microscopes have integral camera systems that offer a significant advantage in the ability to control the light path through the microscope. They also have an advanced exposure-measurement system that reduces reciprocity and exposure errors. At first, digital cameras did not have sufficient sensitivity for photomicrography, but that is no longer the case. The advantage of using a digital camera, especially one wired directly to a computer, is that the investigator can make several adjustments to get good images while the sample is mounted in the microscope.

FISH-EYE PHOTOGRAPHY

This is the process of capturing images using a fish-eye lens. Fish-eye lenses are very wide-angle lenses that take in an almost hemispherical image. The focal lengths of fish-eye lenses depend on the size of the format of the image frame (e.g., 35 mm, 120, digital, etc.). For 35 mm cameras, typical focal lengths of fish-eye lenses are between 8 mm and 10 mm for circular lenses, and 15 mm and 16 mm for full-frame lenses. All ultra wide-angle lenses can suffer from some amount of distortion, but the fact is that the elements to the sides are physically further away from the camera than the elements near the center of the object frame. As a result, the items to the sides are properly rendered as smaller. Fish-eye lenses achieve extremely wide angles of view by foregoing a rectilinear image. They opt instead for a special mapping, which gives these images their characteristic convex appearance.

Fish-eye lenses often are used by photographers shooting broad landscapes to suggest the curve of the Earth. Hemispherical photography sometimes is used for various scientific purposes to study plant canopy geometry and to calculate near-ground solar radiation. There are tools available in image-editing software packages that allow the photographer to do some corrections for curvature associated with short focal length photography. The objective, however, is not necessarily to render the image as a person would see it, but to get it all in one image. Film and digital cameras can be used equally as well with these lenses.

TELEPHOTOGRAPHY

This is a process or technique for photographing distant objects and making them appear as if they were close. It depends on a telephoto lens. The distant objects will appear magnified. A telephoto lens is a specific construction of a long focal length photographic lens that places its optical center outside of its physical construction, such that the entire lens assembly is between the optical center and the focal plane. A lens with a focal length longer than a normal lens is not necessarily considered a telephoto lens. A telephoto lens must incorporate a special lens group known as a *telephoto group*. But most of the time, lenses with a long focal length are referred to as telephoto lenses. Telephoto lenses have a smaller angle of view than normal lenses and compress the image (making the background appear closer to the foreground). These lenses can be used with either film or digital cameras.

EXPOSURE COMPENSATION

This is a technique used to employ an exposure value that is different from what we might expect, based on a metering of the subject scene. It is sometimes referred to as EV (exposure value). There are many factors that may cause the indicated exposure to result in a less-than-optimal image, including unusual lighting, use of filters, nonstandard processing, or purposely underexposing or overexposing for purposes of emphasis. Some cameras include exposure compensation as a feature to allow the user to easily override and adjust relative to the automatically calculated exposure. The compensation can be either positive or negative, and is commonly available in third- or half-stop increments.

Camera exposure compensation is commonly stated in terms of exposure value—1 EV, which is equal to one exposure step. An example of using EV could include a dark subject against a light background, or a light subject against a dark background. This technique works equally well with both film and digital photography, although it can be more important with slide films or digital cameras where there is less dynamic range.

SUMMARY

In general, the techniques developed by film photographers are applicable to use with digital cameras. There are some techniques that are essentially useful only in the digital domain—for example, the three-dimensional scanning. And since there is very little new design work on film cameras, we can expect that the new techniques will be primarily digital.

CHAPTER 10

Image Processing Tools

One of the biggest advantages afforded by digital imaging technology is the ability to quickly and easily employ a range of image-processing tools. Almost every tool currently in practice is mathematically derived from previously known dark-room techniques. It took a real expert and a lot of time to use the dark-room techniques well, but they indeed were used a lot over many years. Films like *Jurassic Park* brought digital technology forcefully into being in the movie (photographic) world. But before that, *It's a Mad, Mad, Mad, Mad World* convincingly employed amazing trickery without the benefit of computers. With digital-image processing, the approach can be applied rapidly, examined, and redone as many times as necessary to get the desired effect. There is no need to go to a dark room and complete a wet process in order to check out the results of each iteration. This is the great advantage in the commercial world, and the great concern in the forensic world. Over the last decade, a team of experts has been working to ensure that forensic examiners can enjoy the advantage of easy processing while avoiding the concerns of manipulation. This group, the Scientific Working Group on Imaging Technology, was started by the Federal Bureau of Investigation in 1997 and comprises a range of people from all over the United States, and a few from other countries. Their findings are summarized in Chapter 18.

The forensic community wishes to avoid changes that lead to a false or misleading interpretation of the content of an image. For example, assume there is a crime scene photo that shows a TV remote under a table and it is important to know whether its brand is SONY. Many tools and techniques could be used to enhance the image to show that it says SONY, says something else, or is just not sufficiently clear for interpretation. The exact set of tools and their order is not necessarily important to know; and if two or three processes are used, they are not expected to result in mathematically identical final images. What is expected is the ability to speak with authority to the question of SONY or not, or don't know. Likewise, in a fingerprint image, what may be of importance is the ridge count between a set of miniature points. Different enhancement processes may lead to images that are not mathematically identical, but if they are all legitimately derived from an original image and lead to the same counts, then the content is properly extracted.

Notice in the two scenarios just described it is virtually impossible to imagine a process that will systematically lead to a false reading that implicates a particular individual. That is, the name SONY will not come out to be Panasonic and the ridge counts will not come out to be one specific person instead of another. Rather, the much more likely outcome will be an inconclusive result. Scientifically, an argument that the processing mistakenly included a particular person would have to show how such a thing could be plausible based on the original and the processing steps, or would have to show deceit. The probability of a false inclusion is vanishingly low ("Probability of false positive with an innocent image processing routine," by Dr. Fang Li, Department of Mathematical Sciences, Indiana University Purdue University Indianapolis; Indianapolis, Indiana; submitted for publication to the Journal of Forensic Identification, 2007).

This is not to say that the investigator can be cavalier in processing images. Quite the contrary: there is much to be learned from images if they are carefully and correctly processed. There are two types of errors to consider, one intentional and the other unintentional. The image-processing tools can be quite complicated and, if improperly used, can easily end in a null result. Thus some basic understanding of the tools will help avoid unintentional errors. As for intentional errors, this is not a technology problem. There are virtually no new tricks hidden in digital imaging technology. The need for honesty and integrity permeates all of what is done in a crime laboratory—ensuring this is a well-known management issue.

There are hundreds of image-processing tools in use today and some investigators will conceive and write their own in order to help with an investigation. So given the scope of this book, a few of the more commonly used tools will be described with the intention of helping users understand just what they are doing to an image when they employ that particular tool. The result will be to make it clear that some sequences should be avoided.

Prior to discussing mathematical analytical processes, it is worthwhile to review some basic mathematical concepts that apply. Image processing normally is done using integer arithmetic and the histogram is a common way to evaluate an image.

The tools will be discussed in groups: tone scale adjustments, color adjustments, and image structure adjustments. There will also be a section that deals with tools that are applied to part of an image and not to the whole thing. These are local tools as opposed to global tools. When local adjustments are employed it is good practice not to conceal that fact from the viewer.

MATHEMATICAL CONCEPTS

In general, images are processed using integer number scales. The scales are of limited range and never go below zero. Integers are whole numbers—numbers cannot have decimals like 240.4 or 240.6, only 240 or 241, respectively would be allowed. So, the average of 240.4 and 240.6 will be 241 if a round-off technique is employed, and 240 if truncation is used. The real average (using *real*, as opposed to *integer*) is 240.5. If this is rounded off, it will convert the result to 241 since no fractional component is allowed and since the implied fractional component is one half (or greater). Note this is 241 and not 241.0 or some other value that implies a fractional component. In the case of truncation, any implied fractional component is simply dropped. The result is that the average is 240. Clearly, one part in 240, or 0.4%, is rather small, but the same rules would apply if the initial values were 2.4 and 2.6. Now the result (truncation) would be 2, or 20% error. What this means is that any time division is applied, there is a chance for error. This also holds for functions like taking a square root, a logarithm, using a trig transformation, and so on. In general, applying arithmetic to integer numbers can lead to errors and related strange behaviors.

Another consideration is that unless we take efforts to the contrary, the range of allowable numbers is limited. The most common imaging-related scale is the one of integer numbers from 0 (zero) to 255. No negative numbers are allowed and no numbers from 256 and higher will be allowed. In this case, if we were increasing the contrast of an image, we would multiply each pixel value by a constant and the result would be the new pixel value. So for example, take a value of 135 and multiply it by 2. We might expect a new value of 270, but this number is not allowed. The arithmetic will show 255 instead—the biggest allowed number. Taking the example further, if the pixel with a value of 135 were next to a pixel with a value of 140, the second one would appear a bit brighter than the first. But, like the first, when the multiplier is applied, the new value will still be 255. And once the calculations are complete it will be impossible to know that originally these two pixels had different values. The original information is irrevocably lost.

Similar effects will happen at the low-number range of the scale, where there is a tendency to have a mathematical operation that normally would result in a negative number. But since negative numbers are not allowed, a zero will be returned instead. Again, the original information will be irrevocably lost. The additional complication of large percent errors due to truncation at the low-number end of the scale can exacerbate these limit problems.

Beyond the range and truncation limitations there is a sampling issue. An object in a scene can have a brightness value of any level. The sensor chip in the camera or scanner starts out as an analog device and will represent all these fine values. But in order to make a digital image, the analog brightness level will be converted to a digital value. But only the integer values in the allowed range can be assigned. The result is that some information will be lost immediately as the analog-to-digital (atd) conversion is made. These errors will usually be of little consequence since the human visual system will probably not be sensitive to the small nature of the errors. If, however, we wanted to analyze the image by exaggerating these differences, it would not be possible once the conversion had been made.

One way to reduce these effects is to use a number scale with more numbers. The 0–255 scale just described is the one most frequently used in image processing and display because it is generally agreed that humans cannot distinguish gradations finer than 1 in 256. Also, 256, when represented as a binary number, comes out to be 2^8, or one binary byte of 8 bits. This is very convenient for computers; but we can do better if we use 2^{12} or 2^{16}, as many digital cameras do. The 12-bit scale has 4096 integers and the 16-bit scale has 65,536 integers. Note that in all three scales, the darkest value will be zero, and the brightest value will be he highest number allowed—that is, 255, 4095, and 65,535, respectively. And when it comes time to create visible images (printed or projected), the scale will be converted back to the 8-bit scale since that is all that can be seen anyway. Clearly with scales that have more bits, truncation errors will have smaller implications. However, the range limitations (restricting the biggest and smallest numbers allowed) will still apply.

The implication is that certain steps should precede others if a sequence of image-enhancement tools is to be applied. To see an example, look at the images in Figure 10.1. Image A is the original image and image B is that image after increasing brightness and contrast. Presumably the brightness and contrast increments were made for other parts of the image not shown in these outtakes. Note that if we were to try to make measurements of the characters in the images, different values would be found. It would not be sound science to try to use the processed image to try to identify a typewriter after using significant amounts of brightness and contrast adjustment, even though brightness and contrast are considered relatively safe enhancement tools.

One common tool for examining tonal content (the levels of brightness and darkness in an image) is to create a histogram. A tonal histogram is a graph in which the horizontal axis denotes the levels (typically 0–255) and the vertical axis shows how many pixels in the image are at that level. For example, consider an image of a perfectly uniform gray patch with one million pixels, in which all the pixels are at 128. The resulting histogram would be a graph with a single tall column located at 128 on the horizontal axis and it would have a height of one million. If there were four equal area patches—one at 60, one at 120, one at 180, and the rest at 255—the graph would have four columns, each with a height 250,000 pixels and located at 60, 120, 180, and 255 on the horizontal axis. As the image gets more and more complex, the general shape will diverge from that of simple tall columns and appear as a continuous shape. It will peak at the point on the horizontal axis that represents the most common pixel value.

A well-taken photo will have some number of pixels at each level on the horizontal axis (see Figure 10.2). If they are all jammed up at the dark end of the

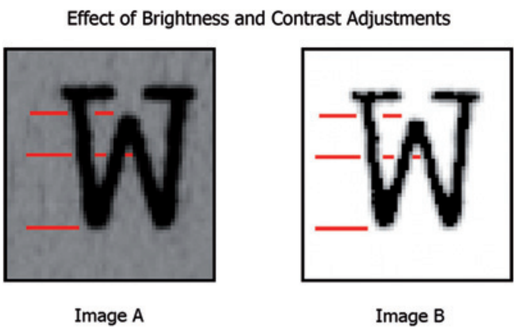


FIGURE 10.1 *The effect of applying significant amounts of brightness and contrast adjustment are shown. In addition to making the image lighter and imposing more contrast, the tools have made the lines thinner, and the dimensions different. The red lines retain the same spacing but they no longer meet the image in the same places.*

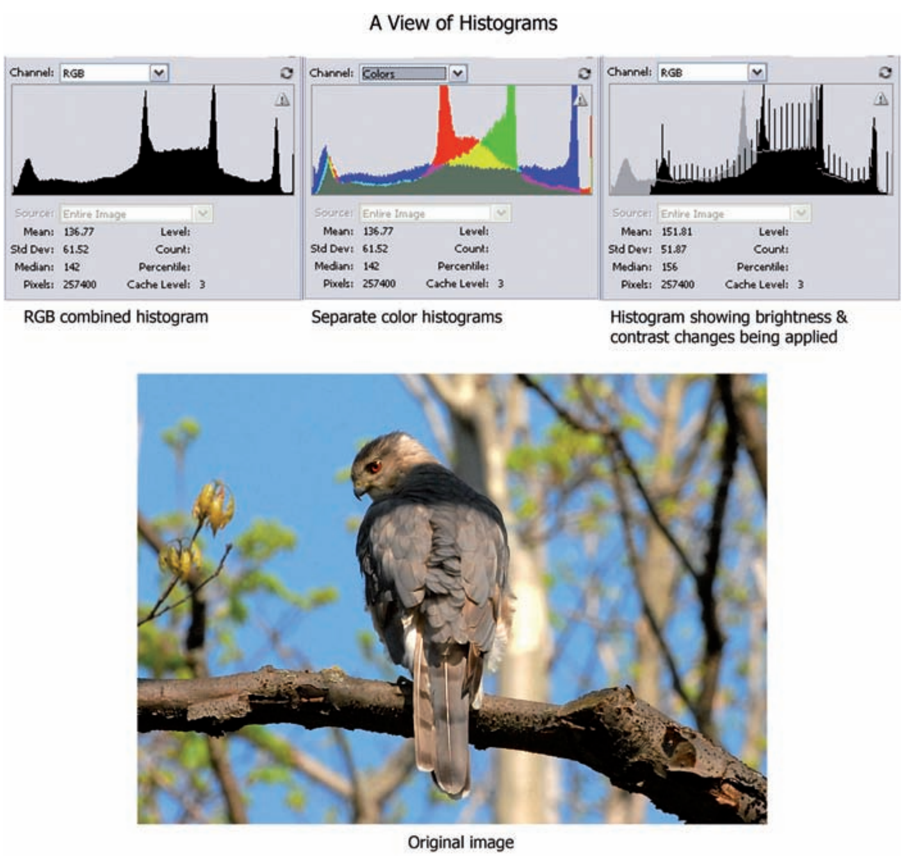


FIGURE 10.2 *Different types of histograms are shown. The original image is displayed along with the histogram that shows the combined colors and then the separate colors. The third histogram shows the gaps created when the brightness and contrast are adjusted. Photo compliments of Michael Barth III, printed with permission.*

scale, (see Figure 10.3) the picture was probably underexposed. If they are jammed up at the bright end of the scale (see Figure 10.4) the image was probably overexposed. Many of the newer cameras have a built-in histogram capability and the photographer can view a miniature version of the histogram of a photo just taken, right on the camera's view screen. These sometimes give three histograms on the same graph, one each for the red, green, and blue records.

We can use the histogram to see the effects of scale limitation discussed at the beginning of this section. The histogram at the right in the figure shows the histogram of an original image and then that same image after the application of brightness and contrast adjustments. Note the spikes and gaps. The gaps resulted from truncation errors, causing certain levels to be abandoned, and the spikes show where these pixels accumulated after the calculations.

TONE SCALE ADJUSTMENTS

Brightness Adjustment

Brightness in a reflection print is a measure of the percent of incoming light that is reflected to the viewer. When more light is reflected, the print is said to be brighter. In a projected image, the brightness is an indicator of the amount of light projected. To make an image "brighter," we increase uniformly the amount of light to all parts of the image. The dark portions get more light as well as the bright ones. Clearly if there are areas that are already as light as they can get, they will not get any brighter (the maximum number, 255 in the 8-bit scale). Mathematically this is done by adding a constant to the value of each pixel. If a pixel reading was 50 and the brightness is increased by 35, the new reading for that pixel is 85. A pixel that was 150 becomes 185, and so on. The effect is to shift the sensitometric curve upward as shown in Chapter 2. The equation for a brightness adjustment is simply:

$$Vo = Vi + K$$

where Vo is the output value of a pixel and Vi is the initial value. K is the amount of the increment set by the user. Note that K can be either a positive number or a negative number. Using a negative number makes the image darker. This equation, with the same value for K, is applied to each and every pixel. Figures 10.3 and 10.4 show a dark and a light photograph with the associated histograms.

Following this example, and as was pointed out earlier in this chapter, any pixels with initial values of 220 or more will all end up being 255. All gradations between 220 and 255 will be lost. If a smaller increase in brightness is used, the resulting loss of information will be less.

Typically the user is presented with a graphical interface with a slider that controls K. As the slider is moved, the operator can see the changes in

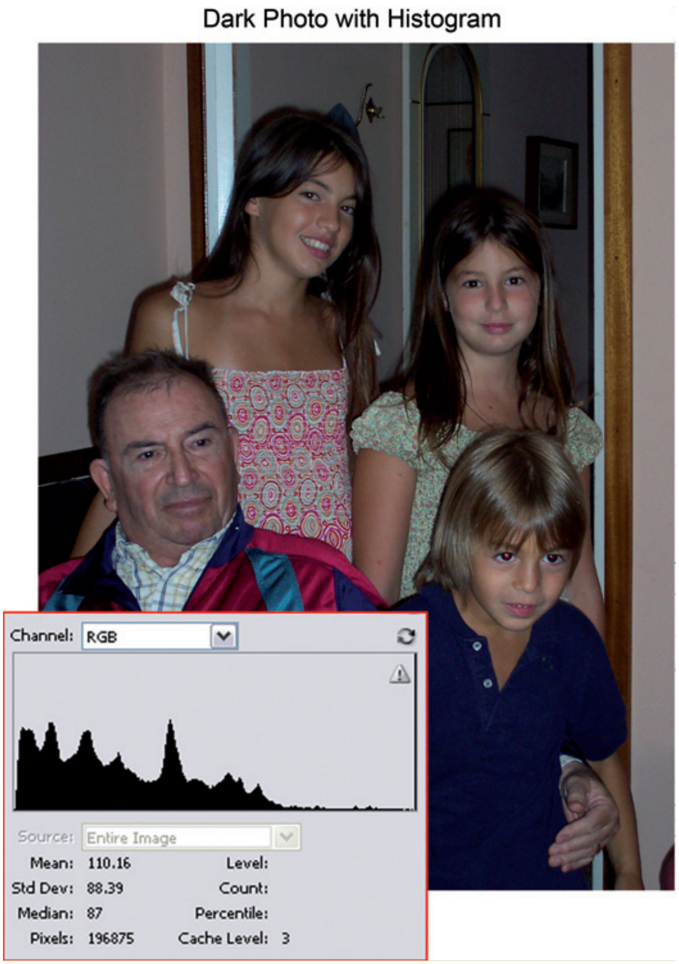


FIGURE 10.3 *Dark Photo with Histogram. Shown is a fairly dark image and its histogram. Note that the majority of the pixels are seen on the left, indicating low brightness values.*

the opened image on the screen. For convenience, numbers are displayed. Moving the slider to the right makes K larger and the image brighter. Moving it to the left makes K greater in magnitude, but into negative values, making the image darker. (see Fig. 10.5).

Contrast Adjustment

Contrast is an indication of how dark the darker patches are compared to the lighter ones. If an image has essentially all its pixels within a small value range, it will tend to look dull. The histogram will show a concentration of the pixels near the center of the horizontal axis, and very few, if any, at the edges. Conversely, if an image shows concentrations at the two ends of the

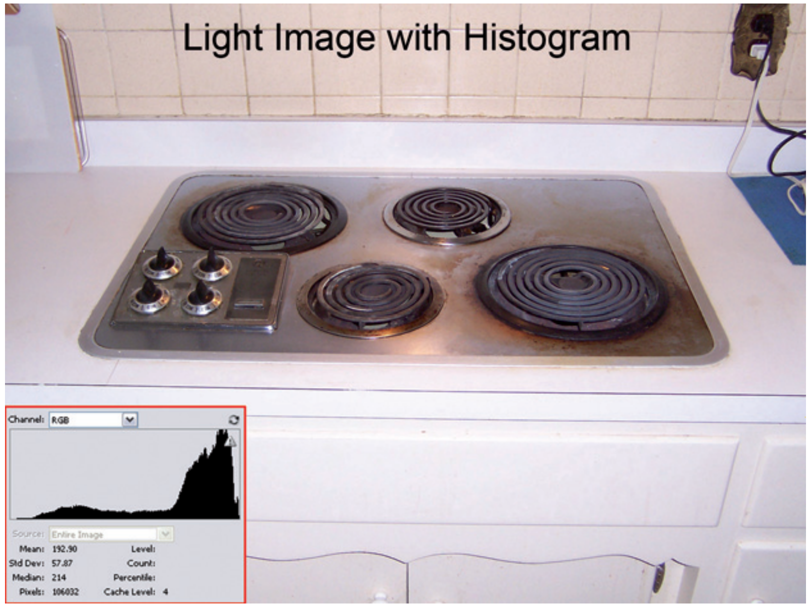


FIGURE 10.4 *Light Photo with Histogram. Shown is a fairly light image and its associated histogram. Note that the majority of the pixels are on the right, indicating high brightness levels.*

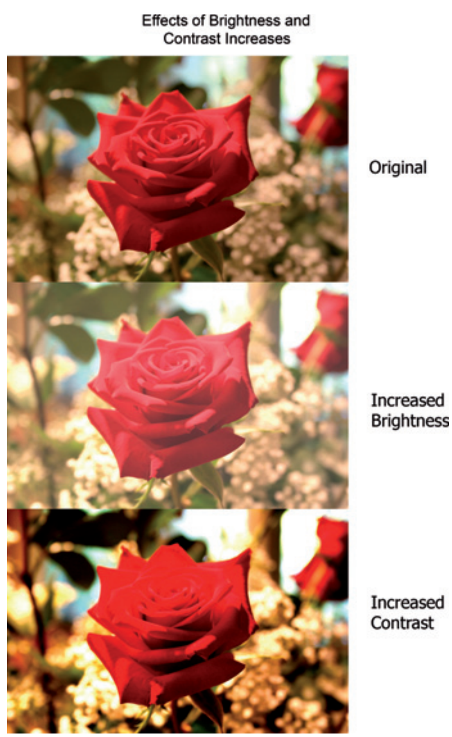


FIGURE 10.5 *Effects of Brightness and Contrast Changes. The top image is the original photo. Below it is shown the same image with the brightness increased. The lower image is the same photo but with the contrast increased.*

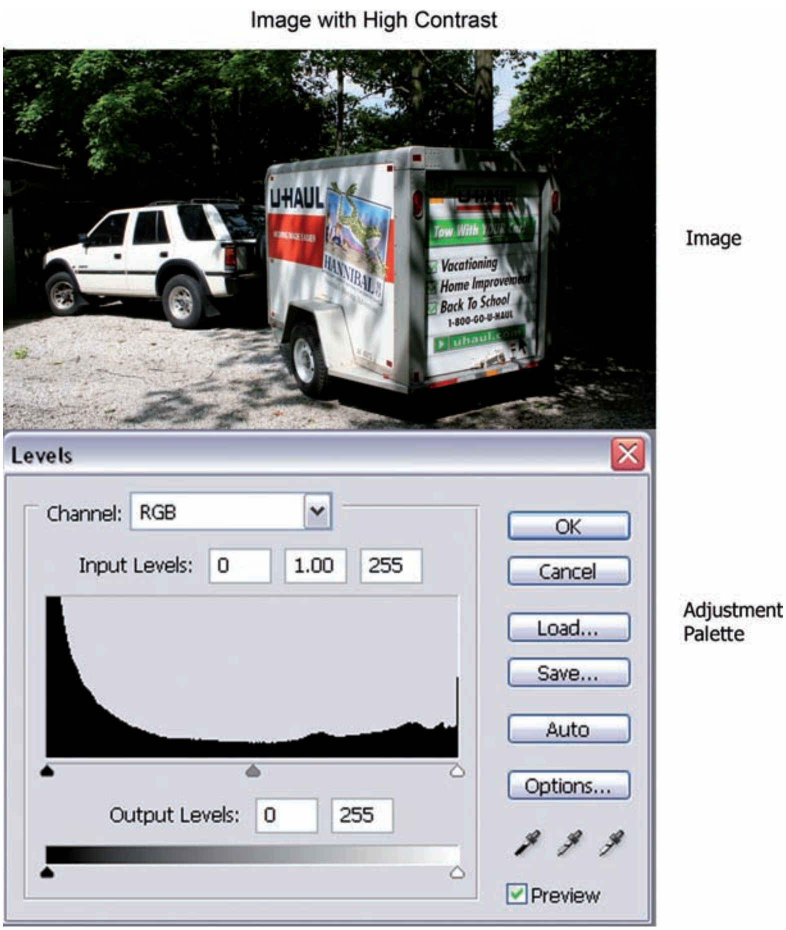


FIGURE 10.6 *Using the Levels Tool. The original photo is quite contrasty. It will be possible to reduce this to some degree by moving the sliders whown in the dialog box.*

horizontal axis and a hollowing-out in the center, the image will be very high in contrast (see Figures 10.6 and 10.7). By comparison, Figure 10.8 shows the same image after histogram equalization.

Mathematically, two steps are used to change the contrast of an image. The first step is to shift the point where the horizontal axis intersects the vertical axis from zero to halfway up the scale. In the 8-bit scale, this will bring the horizontal axis up to 128 (zero to 255 includes 256 steps, and half of 256 is 128). Then each pixel's shifted value is multiplied by a constant. Once the multiplication is completed, the horizontal axis is moved back down to a zero-intersect point again. The equation is as follows:

$$Vo = ((Vi - 128)^{*} M) + 128$$

where Vo and Vi are the output values and input values, respectively, and M is the contrast adjustment factor chosen by the operator. In this version of

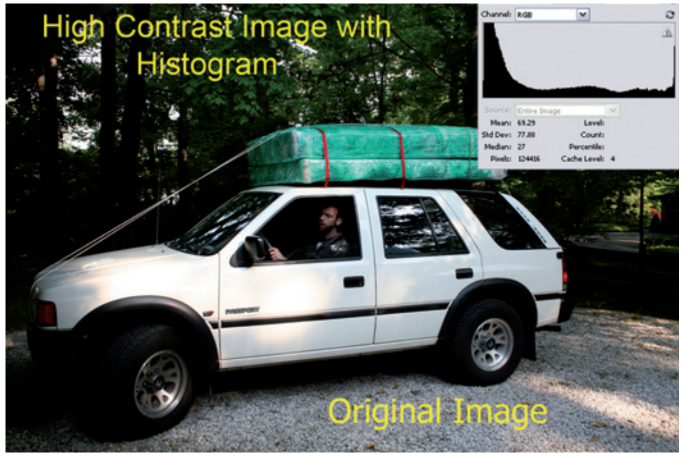


FIGURE 10.7 *High Contrast Photo with Histogram. Shown are a contrasty photo and its associated histogram. Note that there are many pixels on the left and on the right and very few in the center.*

the equation, the 8-bit scale is chosen, hence the value of 128. The shift in the axis causes the sensitometric curve to rotate on its central point.

Note that if M is greater than one, the output image will have higher contrast and if it is less than one, but greater than zero, it will have less contrast. If M has a negative value, then the image will be reversed—that is, dark areas will become light and light areas will become dark, and the image will be transformed from a positive image to a negative image. The common implementations of simple contrast controllers let us increase or decrease contrast; but to change the polarity requires the use of other tools. However, mathematically they are the same.

Notice that all pixels that start off with values of 128 end up with values of 128—they don't change. All others change to some degree. If the initial value was greater than 128 and if $M > 1$, the ending values will be greater than the starting values; and if the initial values were less than 128, they will be lower in value in the end. Thus the darks get darker compared to the lights. If $M < 1$, the darks get less dark and the lights get darker. This has the effect of reducing the contrast, or making the lights and darks more alike. If M is set to zero, all pixels come out at 128—mid-tone gray—and all information in the image is lost.

As was the case with the brightness control, use of the contrast adjuster will result in a loss of information at the extremes due to number range limitations. Consider that the value for M is set to 2. In this case, with the 8-bit scale, all initial pixel values greater than 191 will become 255 (white); and all those with values below 64 will become zero (black). As earlier, any information associated with these groups of pixels will be lost.

A common implementation of this tool is a slider that controls M. Moving the slider to the right increases M above 1 and moving it to the left decreases it below 1 (but not below zero). The operator can see the effects of the changes on the image as the slider is moved. It is also possible to type in a number instead of using the slider. Once the tool is applied and the image saved, the lost information is not retrievable unless a history function is used as well.

Curve Adjustments

Because it is often the case that images do not respond well to uniform changes in brightness and contrast over their entire range, a common tool to use is a curve adjustment tool. This allows the operator to apply some brightness and contrast adjustments to portions of the sensitometric curve while retaining smooth transitions from one portion to adjacent ones.

To use this tool, the operator is presented with a graphical interface. It shows a straight, diagonal line that maps the input (horizontal axis) and output values (vertical axis) for the image. If no action is taken, the output values are equal to the input values. To take action, the operator uses the mouse to move points on the line. Grabbing the line in the middle and moving it upward increases the output brightness of input pixels with values at the middle of the scale. The line bends smoothly, indicating that points distant from the center are changed less and less as they are increasingly far away. Interpreting this change, the central pixels are given a boost in brightness. Points above that level are compressed more closely together on the remaining upper portion of the curve and points below are stretched farther apart from each other on the lower portion. This means that contrast in the upper ranges is reduced and contrast in the darker ranges is increased. The process can be repeated at several points on the curve, adjusting the brightness of selected pixels and changing contrast levels among remaining pixels. The image can be viewed as this is done so a desired degree of change can be employed.

The endpoints of the line will not move unless the operator purposely moves them. This means that there will not be range limitation problems normally associated with brightness and contrast unless the operator purposely chooses to impose these. Leaving the endpoints in place and not pushing the curve to the point where it becomes horizontal and flat in portions means that information is not lost. It may become less visible, but it is not lost.

Equalization Adjustments

More accurately, this is histogram equalization. It refers to a process by which crowded areas of the histogram are spread out and pixels are assigned to other levels. A simple version of this is sometimes referred to as the *levels tool*. This tool provides a graphical user interface in which the image histogram is presented (see Figure 10.8). Sliders (and numerical controls as well) are provided to effectively spread the histogram out and utilize underpopulated brightness levels. If the histogram does not spread all the way to the

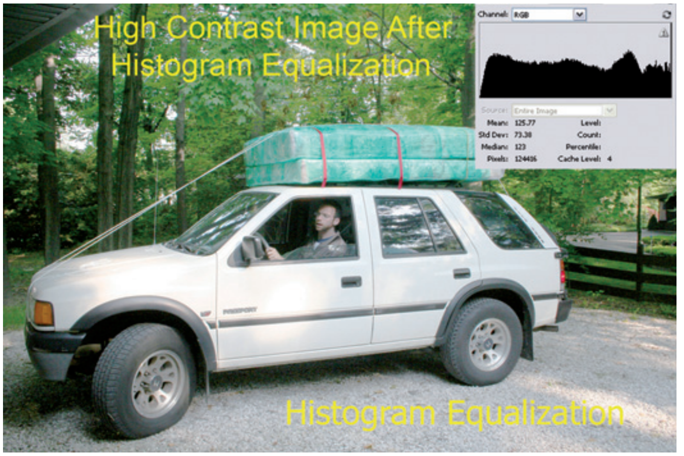


FIGURE 10.8 *Effect of Histogram Equalization. This is the same image as shown in Figure 10.7, but after applying histogram equalization. The pixel values have been spread across the full brightness range.*

darkest levels (near to and including zero), the operator can move the dark point slider to the edge of the histogram on that end, and then repeat the process on the other end. This has the effect of stretching the histogram by forcing the darkest pixels in the image to take on the darkest levels available and then forcing the lightest pixels to assume the brightest levels available. It is also possible to shift the center of the distribution either to take on darker or lighter values. With this tool the operator has the ability to view the image as the sliders are moved.

Another way to accomplish the adjustments is to use the *curves* tool. Assume that the histogram of an image has a peak in the area between levels 120 and 150. Also assume that there are no pixels below 15 and none above 235. Using the curves tool, we could move the endpoints such that the line that shows the mapping of the pixels starts at 15 on the input (horizontal) axis and ends at 235 on this axis. Next the curve can be grabbed just above the 150 point and pulled upward, and then grabbed just below the 120 point and pulled a bit downward. This will stretch the curve between these points and give more output change per unit of input change in this range. The result is to take some of the bulge out of the center of the distribution and to better utilize the full output brightness range.

Both these approaches have the effect of taking pixel values from the areas where there are large groups close together. This means that the original image does not have much contrast in this region. Consider a photo of a marble countertop in which the marble ranges from light gray to slightly darker gray. The contrast of the pattern is rather subdued. If there were a light-to-medium gray liquid spilled on the countertop it would be very hard to see. Stretching the histogram in this area would make the patterns more pronounced and easier to see. The histogram of the image would presumably show a large concentration of pixels with light-to-medium gray values and not much at high and low ends of the range.

There is a mathematical technique that can automatically apply the indicated reapportionment of the pixel values. This is the histogram equalization tool. This tool uses a probabilistic analysis of the distribution of the pixel levels and stretches that portion that shows unusually high and unusually low concentrations. It then calculates a transformation function that allows the original values to be systematically revised to increase the previously small differences. The tool is essentially doing what was done manually with either the curves or the levels tools in the earlier discussions. It can be very effective in expanding the contrast of portions of an image that suffer from inadequate contrast. If there are portions of the image that are not subtle, it can cause some exaggerations that are unacceptable, so it must be tried for effectiveness, but ultimately might not be used. Note that when tone scale tools are used properly, there is no need to lose information, no information is added, and the imagery that is enhanced is purely derived from the content of the original image.

Tonal Scale Gaps

Any of the tone scale adjustments will result in the creation of gaps in the tone scale. This was shown earlier in Figure 10.2, where the histogram of an image before and after application of the levels tool is in the upper right corner. Note that the originally smooth distribution becomes one with gaps and spikes included. These result from the various processes of reassigning pixel values and are aided and abetted by round-off errors induced by the integer scales. These are usually not a serious consideration as they do not adversely affect the ability to interpret what is in the image and they do not represent a material loss of information. They may, however, prevent a perfect reconstruction of the original image. That is, they can be one-way streets. Once taken, there is no going back.

In summary, the brightness and contrast controls—though easy to use and effective—have some associated problems. They apply the same correction factors across the full tonal range and they tend to cause high and/or low brightness pixel information loss. The curves, levels, and histogram equalization tools require a bit more skill to use but they do not apply the same corrections uniformly across the tonal range and can be used in such a way as to avoid or minimize any truncation-imposed information loss. Forensic work is best done with the curves and levels tools, and not with the brightness and contrast tools.

COLOR ADJUSTMENTS

Color images are really composed of (at least) three combined separation images. In the RGB space they correspond to the red, green, and blue image records, or channels. In the CMYK space there are four channels, even though the K record is derived directly from the cyan, magenta, and yellow channels. These all have somewhat intuitive meaning. This intuitive meaning does not extend as easily to other color-space configurations, such as hue saturation lightness (HSL) or, especially, the CIE/Lab space. Nonetheless each of these has three channels as well. In the preceding section, several tone scale adjustment tools were discussed as though—when used—they are just directly applied to a color image. They are not. The operator may see it that way, but the computer is actually processing each of the separate color images one at a time and recombining the result.

Common image-editing software packages allow the operator to separate the image into its different channels. The tool for doing this is called *channels*. The same tool allows the separations to be recombined. Separations are not a new concept. High-budget movies frequently are shot with a color negative film and the resulting images are composed of dyes. But dyes fade over time. So it is common practice to produce separation reels by printing the negative to three black-and-white films through (one each) a red, a green, and a blue filter. This process results in three separation versions of the original movie color negative. The silver records in the separations are extremely stable over time. It is still possible to make color prints from old color movies such as *Gone with the Wind* by printing from the separations.

Color Additions and Subtractions

Color Balancing

Take a series of family pictures and lay them out on the table. In some of the pictures your uncle's face is a bit yellowish, in others a bit more on the green-blue side, and in others still, he looks as you'd expect to see him. Has he been ill? No, the problem is probably with the lights under which the photos were taken. The yellowish ones were probably under low-wattage tungsten lights, the blue-green ones under fluorescent lights, and the normal looking ones were probably shot outdoors in daylight. A good photofinisher will normally use color balancing to make them all look much like the "normal" photos, but some don't bother. (Go somewhere else for film processing.) Most of the new digital cameras have an automatic white balance, which senses the extra yellow or blue light and causes the camera to compensate. However the range of adjustment is sometimes less than you would like. In those cases, chalk up the coloration to artistic atmosphere.

The issue here is how we accomplish color balancing, and this depends upon the color space in which we are working. The easy place to start is with the simple additive RGB space, which actually drags in its relative, the subtractive CMY space. In this combined space, that first image might appear yellowish because the blue channel is too light, or its opposite, the yellow is too bright. Likewise the blue-green image has too much cyan brightness, or not enough of its opposite, red. Color balance is achieved by systematically brightening the colors that appear to be missing or darkening those that are too much in evidence. Figure 10.9 shows a "ring around." In the center is a normally balanced image and surrounding it are six versions of it with one of the additive or subtractive colors brightened.

FIGURE 10.9 *Ring Around. The original image is in the center. Above is one balanced to the red and below to the cyan. Yellow and blue are shown next (clockwise) and green and magenta are next.*

One problem with simply brightening or darkening one of the separations is that it may not be that the whole tonal scale is out of balance. So, just as the curves and levels tools let us alter portions of the tonal scale selectively, most color-balancing tools do the same. This is done by letting the operator apply separate adjustments to the highlights, mid-tones, and shadow portions of the tone scale. It takes some practice, but we can generally achieve a reasonably good color balance using these tools.

It should be pointed out that only so much can be fixed with any balancing tool. The problem usually starts with the light falling on the original scene. For most common light sources, the source itself will appear to be white. If it is off by a little, correction is possible. If it is off by a lot, there is no way to make the image appear to be normally balanced. For example, taking photos in a nightclub that is illuminated primarily with blue lights will probably result in an image that shows that atmosphere. Similarly, if you're in a room with tungsten lights, but bright red walls, it will probably not be possible to get normal balancing in images of items on the floor and close to a wall. These effects occur because the range of adjustment from color to color is greater than the dynamic range of the sensor for the three different sets of pixels in the camera.

To appreciate how this can happen, consider again that all color images really are composed of three separation images. Each separation image is basically a monochrome photo in its own right. However, the camera takes all three images at the same time—using one speed setting, one f/stop and one exposure time. If there is an over-abundance of, let's say, blue light, the blue separation will be overexposed and details in the highlights portion of the image will be bleached out. Attempts to darken this record will not result in reconstruction of the highlight detail since it was never captured in the first place. A camera that can independently capture the three separations would be needed, and no such camera exists commercially.

Hue Lightness Saturation

Sometimes an image just appears to be too gray. The colors are not as vibrant as you might expect. This can happen if the contrast of the original items is low in the area where more vibrancy was expected, or if the lighting was not particularly flattering to the subject. For example, consider bluish light falling on a yellowish subject. One way to deal with this is to use the *hue saturation lightness* (HSL) tool. It can be used to increase the saturation, which refers to the purity of a color. A rich, vibrant, bright red is highly saturated (very pure red). Pink or a grayish pink are desaturated reds. Hue is the dominant color—in this case red. Increasing the saturation of the red will decrease the relative amounts of nonred in the color. Decreasing the saturation of a color will move it closer to gray. Consider that the colors in an image are converted to amounts of hue, saturation and lightness from primary colors red, green and blue. Now multiply all the saturation values in the image by a constant number greater than 1. The resulting image will have higher saturation, and when converted back to red, green and blue values, the change will be noticeable.

Lightness is an indication of how bright an item appears to be. Increasing the lightness is similar to adjusting the brightness of an image—making all objects brighter. However the visual impression of brightness can change the perception of colors. So, increasing the brightness values of the three RGB records by the same amount to get a brighter image will result in a color shift as well. In the HSL space, this is correctly accounted for and the result is the ability to shift brightness without altering how colors are seen. The brightness tool, usually included in a simple brightness or contrast tool set, does the same thing, unbeknownst to the user.

The hue adjustment capability will shift all the colors at the same time. If the reds are moved a bit closer to the yellow, then green will shift the same amount to the cyan and blue will shift accordingly to the magenta. It is like a merry-go-round of color changes. HSL tools are rarely used to change color balance or brightness, but are very useful in adjusting saturation. In a H, S, L rendered image, H is the angle assigned to each hue. Now add a constant to each hue value in the image. This will shift all the colors all at once. When converted back to a RGB image the effect will be quite noticeable.

CIE Positioning

In the 1930s, the Commission Internationale d'Eclairage (CIE) developed a *chromaticity diagram*, which shows all the colors that people can see. It has a shape something like a short, flattened football. It shows black at one end of the ball and white on the other. The straight line going up through the middle of the solid shows various shades of gray. A cross-section through the middle of the football is somewhat like a horseshoe that is closed across the points. It is round on one side and flat on the other. The colors are arrayed around the central vertical line, each hue coming off at a different angle. The farther you go from the gray in the center to the pure colors on the periphery, the more saturated the color.

The modern version of this approach is referred to as CIE/Lab. To locate a color, one needs to know how far up the central line—the "L" dimension—it is (somewhere between black and white); and where it is on the horizontal plane, which has dimensions "a" and "b". The three dimensions of this space are L, a, and b—or Lab. The original CIE space used dimensions labeled Z, x and y. Forty years later these were updated and Z has been replaced by L, and a and b are the new x and y. The changes incorporate mathematical operations used to enhance the trueness of perceptual accuracy.

One of the more important features of this diagram is that if there are two lights, each located at some point in the diagram, all the colors that can be produced by adding various amounts of those two lights will be on the line. The relative amounts of the two lights will depend upon the relative distances from the mixed color to each of the two original lights. If there are three lights, then they will form a triangle, and all the colors that can be produced by those lights will lie within that triangle. A change in position up the center line will result in a change only in lightness, and no shift in hue or saturation. This system is very much akin to how people see colors. The previously described color systems have color coordinates that are based on simple linear transformations from one to the other. The CIE space is not linearly related to the other color spaces.

White Balance

Just over a century ago, Max Planck published a paper that included a remarkable equation. With his equation, the amount of light at each wavelength that is coming from a light source could be estimated by simply knowing the temperature of the source—assuming the source was a "black body." A black body does not reflect any light, but simply emits light as it gets hot. The filament in a standard light bulb is a very good approximation of a black body. Not only is the equation of Planck accurate, but it has only one parameter (and several universal constants): the temperature. The temperature of a black body associated with a specific spectrum is known as its *color temperature*. Since many light sources do not behave exactly as Planck's theoretical black body does, the color temperature of a more common source is known as its *correlated color temperature* (CCT).

We can plot the locus of black bodies at various temperatures on the CIE diagram. The locus forms a shallow arc that reaches from baby blue to pink, and it is peaked slightly toward the green in the center. As the source goes up in temperature, it goes from pink to blue. Some common values are as follows:

Source	CCT (degrees Kelvin)
Low wattage tungsten light bulb	1800
High wattage tungsten light bulb	3200
Noon daylight	5000
North sky	10,000

Figure 10.10 shows the spectra (normalized to peak at 100%) and Figure 10.11 shows the locus of the sources plotted on the CIE diagram. Note that the spectra show that the key difference among the different sources is the red-to-blue ratio. More red indicates a lower CCT and more blue a higher

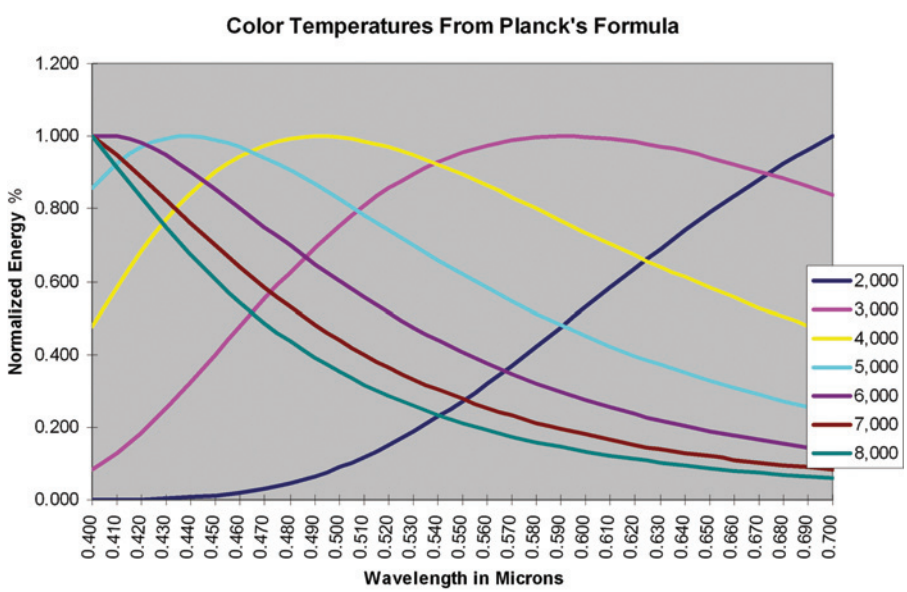


FIGURE 10.10 *Color Temperatures. The various plots show the energy distributions for black body sources at different temperatures. All are normalized to peak at their highest level and calling that 100%. These are directly from Planck's formula.*

CCT. In fact, a reasonable estimate of the red-to-blue ratio can be used to estimate the specific CCT of a given source.

The typical photo has a highly reflective portion somewhere within it, and most of the light reflected off the object will be from this bright section. This is what the human eye normally uses to determine what is white. When people view paper prints in some ambient lighting, they take reference from some other things in the field of view that they know to be white, and mentally adjust the image accordingly. When looking at projected images in a dark room, the cues are taken from the image itself, and it has been demonstrated many times that the coloration of the image can be changed quite a bit before the people looking at it will be bothered by the lack of basis for color balance.

With traditional film photography, the photographer makes an estimate based on knowledge of the source and can place an adjustment filter over the lens before taking the photo. The filtration compensates for any extra red or blue light and makes whites come out white. Digital cameras can do the same. They integrate the light reflecting from the object and into the camera and determine the red-to-blue ratio. This indicates a CCT, and the

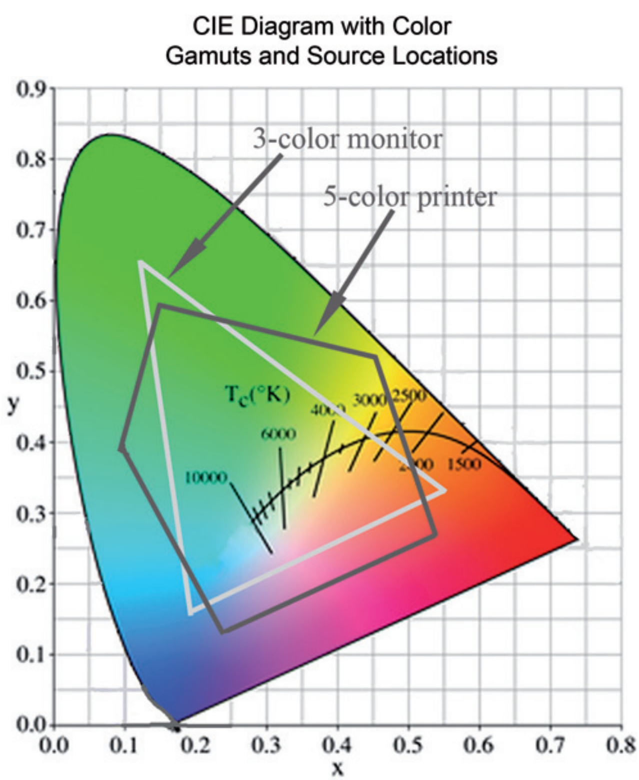


FIGURE 10.11 *The CIE Color Space. Figure 10.10 showed the distributions of different light sources. Here the locus of their locations in the Lab space is shown.*

camera can adjust the response of the camera to red, green, and blue light to make the bright portion appear white in the resulting image. The process is called *automatic white balancing* and it generally works quite well. If there is nothing in the image that is really white, the system will be fooled. For example, a camera taking a picture of a pale pink rose in front of a black background will try to make the rose come out pretty much white. But for most objects, the process works quite well. The forensic photographer takes photos of some rather odd objects and care must be used not to let the camera distort the colorization.

With newer cameras that have a RAW file output capability, the operator can override the camera's estimate of CCT. If the image does not appear well balanced (the camera did not make a good estimate of the CCT), it is possible to type a known value for the source. This will instantly rebalance the image, generally making a dramatic improvement.

IMAGE STRUCTURE ADJUSTMENTS

Interpolation

Many digital cameras will bring the image in at 72 pixels per inch (ppi). This is convenient for viewing on a computer screen, but not for printing, where we want about 300 ppi. Conversely, a scanner might bring an image in at 1000 ppi, and we still would want about 300 ppi for printing. The process of converting an image from one ppi value to another is called *interpolation*. In the former case, the process is called interpolating up, and in the latter it is interpolating down. There are three very common algorithms used for accomplishing this: nearest neighbor, bilinear, and bicubic. For simplicity we will discuss interpolating up by a factor of two in each dimension. This requires estimating new pixel values. Interpolating down, or to other ratios, uses the same basic concepts.

Figure 10.12 shows an extreme close-up of a portion of a gray-scale image. The pixels are clearly visible. To interpolate up by a factor of two in each direction, we need to create space for new pixels. Since we will interpolate up by a factor of two, one space is inserted between each of the adjacent original pixels along the rows and columns, and then one on the diagonal. Next, values must be estimated for each of the new pixels. Following the nearest neighbor approach, the new pixels immediately adjacent to a given original pixel are given the same value as the original pixel. The result is that the image now contains four times as many pixels (two times the number in both dimensions). However, the image looks exactly the same.

The nearest neighbor approach is useful when the ppi needs to be changed, but no change in the image is to be allowed. Presentation packages, such as Microsoft's PowerPoint software, interpolate (resample) images that are inserted into presentations to be properly suited to display devices. It alters the content of an image during this process (a bicubic process,

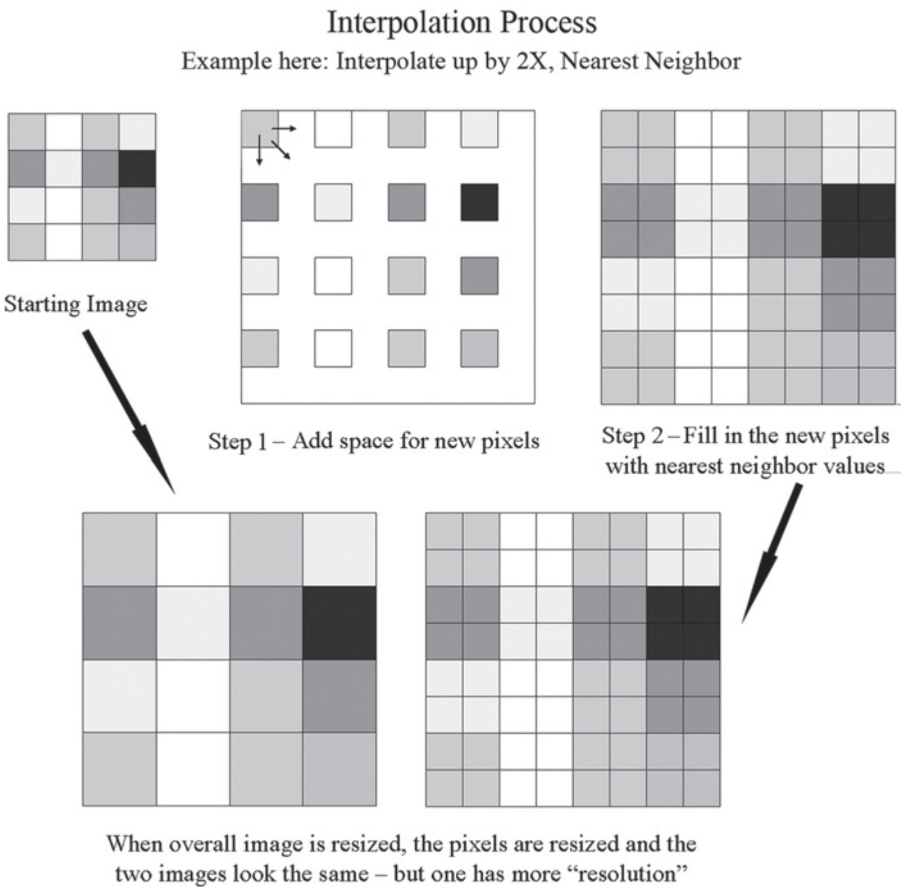


FIGURE 10.12 *Interpolation. The starting image is shown in the upper left corner. The pixels are separated to allow pixel-sized spaces in between each one. New pixels will be placed into each blank space. In this case, the nearest neighbor approach is being used so the new pixels get their brightness levels directly from above and to the left. This results in four times as many pixels, so the image is larger, but it looks the same. Other estimation processes can be used that will change the appearance of the image.*

which we will take up later). It tends to smooth sharp edges. To avoid the alterations due to resampling, the operator can change the ppi to 92 (typical screen values) using nearest neighbor. Then when the image in inserted into PowerPoint, its appearance will not change. In addition, prior to averaging several images together to try to increase the overall resolution, interpolate up using nearest neighbor. In this process, the different image will not have the pixels aligned with the image detail in exactly the same way, and when averaged together, the new image will be able to combine the information from near pixels. But if more detail is to be seen, more ppi are needed. Nearest neighbor retains the pixel information in each image while allowing the averaging to occur at higher resolution.

Although the nearest neighbor approach has its applications, it does not deal with image information in a way that always makes appealing prints. We often want to increase the ppi and smooth the transition between adjacent pixels. This reduces the visibility of the pixels themselves in the print. A simple way to do this is to use the bilinear (bidirectional linear estimation) approach. In this case, the process starts the same way as before—a pixel space is inserted between each original pixel. It differs in how the system determines the value to assign to each of the new pixels. In the bilinear approach, the value assigned to each new pixel is the simple average of the original pixels on either side of the new one. The assumption is that if original pixel #1 has a value of 10, and original pixel #2 has a value of 14, it is reasonable to assume that a new pixel inserted in between them should have a value partway in between. A simple average is as reasonable an estimate as any, it would seem. This approach gives smooth-looking prints and generally works well as long as there are no dramatic changes along pixel rows and columns.

When there may be dramatic differences, another approach can be used. This is the bicubic approach. This one is a bit more complicated, but has ancient roots. The process is based on a cubic spline. The cubic spline method is the same as the French curve method used in the days of hand-drawn mechanical drawings. Basically, a cubic polynomial is estimated for four points—the original pixels in a row.

$$V_o = a \times V^3 + b \times V^2 + \mathrm{c}V^1 + d$$

where V_o is a new value computed from the original values, V, raised to the different powers. The values for a, b, c, and d are estimated using the values for the first four pixels. This is for the first set.

Using the cubic equation, the value for the new, intermediary pixel between the first two pixels is estimated. When this is done, the next three pixels are used, along with the slope of the end of the first cubic equation, and a new cubic equation is estimated. This is used to estimate the next new pixel's value. Then the process keeps moving along. When all the rows are done, the same process is repeated for the columns (*bidirectional cubic spline*). With the new intermediary pixels in the rows and columns filled in, the same process can be used to estimate the new pixels that are on the diagonals. The bicubic approach gives excellent image renditions and is probably the most widely used approach for preparing images for printing.

Both the bilinear and bicubic approaches create values for inserted pixels that were not in the original image. Experience shows that interpolating up by a factor of two, or as much as four, does not misrepresent the original image. Interpolating up by more than this is not recommended since too many of the pixels in the final image are the result of arithmetic calculation as opposed to the performance of the camera.

Note that the bilinear approach can estimate only straight lines between pixels, so more intricate patterns might be missed. The bicubic approach can interpolate using curves and is more sensitive to details.

Sharpening

The standard sharpening process is based on the use of a *kernel*. This is a small matrix of numbers, usually nine, arrayed in a square. The kernel can be thought to superimpose on the first square of nine pixels in the upper left corner of the original image. The value of each of the underlying pixels is multiplied by the numbers in the corresponding kernel cell locations and the resulting products are added together. The resulting number is placed in the center of the kernel placement in a corresponding new, sharpened image. Then the kernel is moved along to the next set of nine pixels, dropping the first column and adding the fourth. This process is repeated across the entire image in both directions, creating a new image in its wake.

The basis of the process is that the value in the center of the kernel is equal to one plus the values of the surrounding eight locations. So, as shown in Figure 10.13, the center value is nine and the values in each of the surrounding eight locations is negative one. Adding the eight negative ones to the positive nine results in a value of one. So, on a portion of an image in which all the covered pixels have the same value, the kernel process returns a value of one times the value of the original image. In other words, it does not change anything. If the kernel is atop a location where the three values in the right-most column are higher than those of the remaining locations, the kernel will return a value that is greater than one times the average value of the nine locations. When the center of the kernel and the right-most column are above pixels with higher values than those below the left-most column of the kernel, it will return a value that is much greater than one times the average. The reverse happens when the kernel encounters a location with lower values to the right. When the process is completed, the computed image will show accentuated changes in value among adjacent pixels. That is, the image will appear to be sharper.

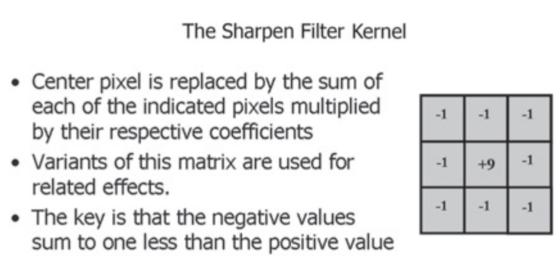


FIGURE 10.13 *The Sharpening Filter. The kernel of the sharpening filter is shown. It is used to compute new values for the pixel under the central spot. Then the kernel is moved and the process repeated. All but the central multiplier is negative and the sum of all the multipliers is equal to 1.0.*

This simple sharpening tool can do a good job on scans of documents, but it can be a bit harsh on continuous tone images, such as crime scene photos. It produces a result similar to the Mackey lines that can be produced in a traditional silver halide image when the film or print stock is processed without agitation.

Blurring

To sharpen an image, a kernel was used that accentuates the differences between adjacent pixels. Blurring, which is essentially unsharpening, uses a kernel that samples nearby pixels and mixes their values in with the pixel being adjusted, partially averaging them, or blending them all together. The most common blurring tool is the Gaussian blur. In this case, the values for the coefficients in the kernel follow a Gaussian distribution. The farther away they are from the center number, the lower they are, and the rate of drop-off is per the normal bell-shaped curve of Gauss. This can be seen in Figure 10.14. The radius is the analog of the standard deviation and the bigger the radius, the more the blurring effect. As with the sharpening kernel, the values of all the numbers in it must sum to one.

Blurring is not used that often in forensic work. Usually the objective is to seek higher resolution, not lower. It can be used selectively to hide an uninvolved person's face or some other such application. It is also a constituent part of the unsharp mask process.

Fourier Transformation

Imagine that we have a photo of an object that is nothing more than a series of vertical lines spread across the frame. Assume that in this case there are 20 lines spread across a one-inch sensor chip. Now let's move the camera back away from the object and take another picture. This one is very similar to the first except that now there are 40 lines across the same one-inch frame. The first image has 20 lines per inch and the second has 40. The term

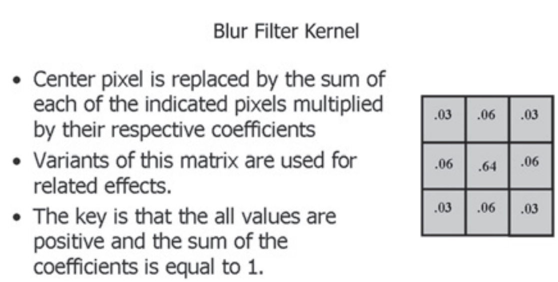


FIGURE 10.14 *The Blur Filter. The kernel of the blur filter is shown. It is used to compute new values for the pixel under the central spot. Then the kernel is moved and the process repeated. All the multipliers are positive and the sum of all the multipliers is equal to 1.0.*

of art is to say that the second image has a higher *spatial* frequency. This is virtually the same as in sound. Middle C has 256 wave peaks per second, and the A above middle C has 440 and has a higher *temporal* frequency. You might tune your FM radio to some favorite station and find that it broadcasts at 100 million Hertz, or cycles per second. Your second favorite station is at 90 megahertz. Your favorite has a higher temporal frequency. The media are different in these examples, but to a large extent, the mathematics involved are the same. In imaging, the main measure is spatial frequency and it will be designated in different ways depending upon the case.

Joseph Fourier, back some 200 years ago, showed that any repeating pattern can be represented as a sum of sine and cosine waves of specific frequencies. The amplitudes and the phases must be determined for each case, but the arbitrary original pattern can be replaced by a mathematically convenient sum of easily processed functions. In the case where a pattern does not repeat by itself, it can be replaced by a new one in which the pattern as a whole is repeated over and again. The process of converting an original pattern to a sum of specific sine and cosine waves is known as Fourier Transformation. Estimating the actual amplitudes and phases for each of the sine and cosine waves is computationally intensive, but modern computers can do it fairly readily using an algorithm known as the Fast Fourier Transform, or FFT. FFT has become the term used to indicate that a Fourier analysis process is being applied. It can also be done using analog computation.

The old radios used to use analog methods to apply the essence of FFT to incoming radio signals. In this way, the station of interest can be sorted out from all the others that reach the radio at the same time. In imaging, the FFT can be used to separate a particular repeating pattern from numerous others—just as the radio accepts one station and rejects all others. The classic application is in fingerprint work. The ridges of the fingerprint are all spaced from each other by approximately the same distance. That is to say they represent a specific spatial frequency. If the print is on a busy background such as a newsprint photo, the dot pattern of the photo will probably have a different main frequency. Note that in image applications, the FFT is applied in both the vertical and horizontal directions. Now just as the radio tuner can tune in one frequency and tune out all others, the FFT for imaging applications can do the same.

Figure 10.15 shows a fingerprint on newsprint and how it might be processed using an FFT software module. The section at the top shows how the image converted to frequency space by application of the FFT. The spot at the center of the plot shows where the darkness of the area is proportional to the amplitude of the sine waves at zero frequency. In essence, this is the average for the whole image. Moving out in any direction from the center, we can see the amplitude associated with increasing frequencies—the farther from the center, the higher the frequency. There is a dark gray ring around the center

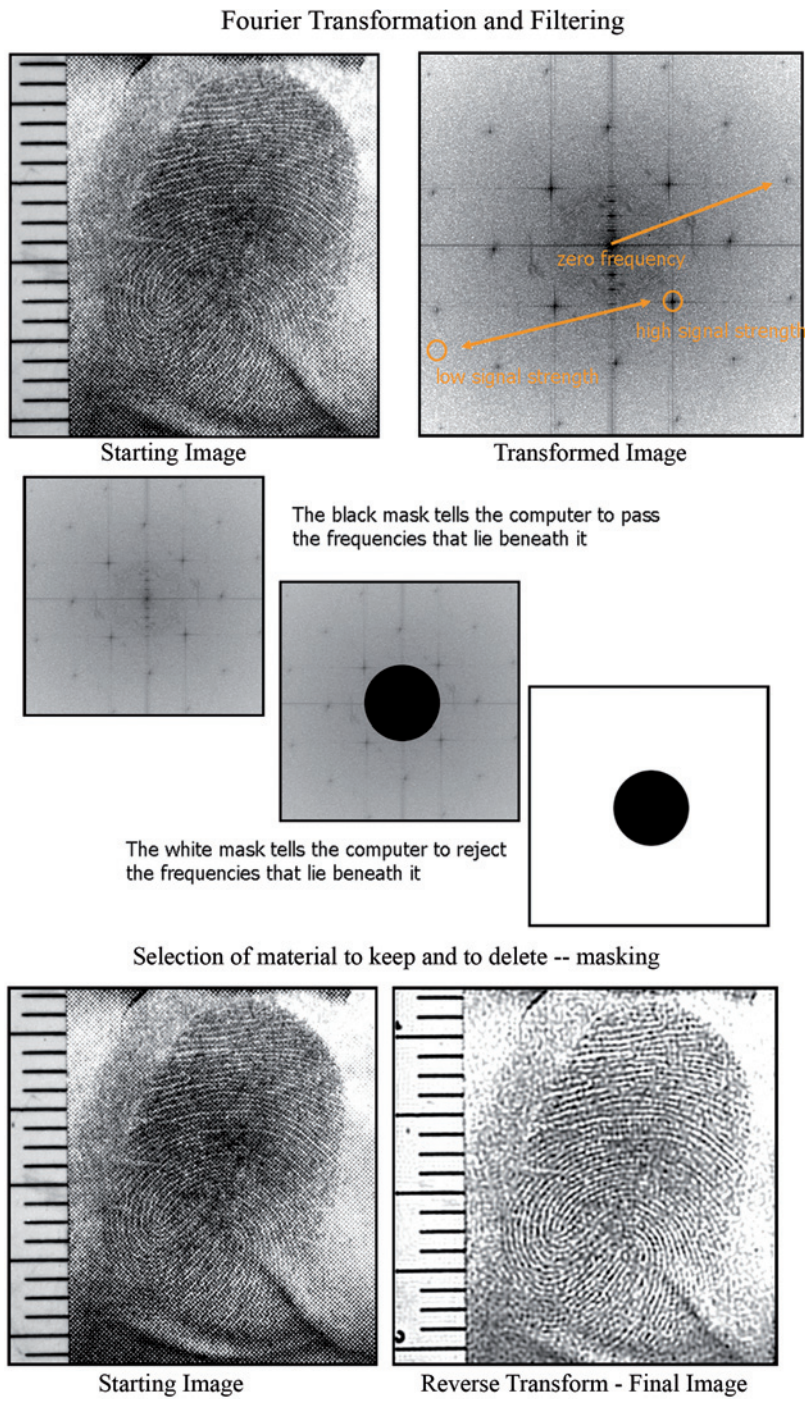


FIGURE 10.15 *Fourier Transformation Filtering. The original image is shown in the upper left. It is transformed to frequency space and areas associated with the unwanted background are discarded and the rest is kept. Then the reverse transform is performed to yield the key information with the background noise reduced. The Fourier Transformation is used widely in many applications.*

spot that represents the information from the fingerprint. Most prints have ridges at all angles, so the result is an annular ring. The isolated spots regularly arrayed elsewhere in the transformed image are due to the newsprint dots. We are now in a position to select information. The fingerprint information is to be kept (the station to which we wish to tune our radio), and all the other information is to be eliminated or reduced (all the stations we don't want to listen to). To accomplish this in the example tool, the information areas where we wish to keep the information are covered by a black mask and the areas to be cleared are covered with white. When this is done, the reverse transform can be applied. The information to be kept will be reproduced, and the information selected for removal will be either eliminated or greatly suppressed. This is shown at the bottom of Figure 10.15. Note that if the background pattern has a spatial frequency close to that of the fingerprint, separation will not be possible.

Unsharp Masking

Now that concepts of sharpening, blurring, and frequency are clear, we can put some of them together to understand the complex process known as unsharp masking. Despite its name, unsharp masking is a process used to make an image appear sharper. Almost all images contain details that are made of sharp demarcations and some that are gradual. The sharp areas contain high-frequency information and the gradual areas are comprised of low-frequency information. To sharpen an image with the unsharp mask, the low-frequency information that is present in the high-frequency areas is reduced and the remaining image contrast is adjusted back up to highlight the high-frequency information that was not suppressed.

The first step in applying the unsharp mask process is to create the mask itself. To do this, a duplicate of the original image is formed and then Gaussian blur is applied. The result is a duplicate of the original from which the high-frequency information has been significantly reduced. Next the image is converted to a negative. This is the unsharp mask. It can be added to the original image to create a third image in which the value of each pixel of the original is added to the value of its counterpart in the mask. Since the mask is a negative, this really amounts to a subtraction. In this way, the low-frequency information of the mask is subtracted from the corresponding information of the original, thus reducing the amount of low-frequency amplitude in the final image. And, since the mask contains no high-frequency information, this aspect of the original is left intact. The result is that the high-frequency information is represented at higher contrast than the low-frequency information. The image contrast is then boosted to better represent the original subject.

The process involves a number of delicate balances. The contrast of the mask must be adjusted down a bit so as to not totally eliminate the low-frequency information, but just to suppress it a bit. The amount of blurring must be adjusted by setting the radius value used in the included Gaussian Blur. If the radius is too small, there will be hardly any effect, and if it is too large, the sharpening will be overstated. Finally, the contrast of the final image may need adjustment to bring the high- and low-frequency blend back to that of a normal image. This can be done manually using the underlying tools. Originally it was done in a darkroom with silver halide film (this is where the process was invented). But most image-editing software programs have a nice tool to ease the complications of doing all the steps by hand. Unsharp mask is the most common tool for sharpening images from scanners and digital cameras that need a bit of sharpening. It will not take a grossly out-of-focus photo and make it sharp and clear. It simply enhances the high-frequency information relative to the low-frequency information that was in the original image. It is the most widely used such tool in forensic image processing. There is an "Unsharp Mask" exercise on the website which shows how this process actually works.

LOCAL VS. GLOBAL ADJUSTMENTS

The leading image-editing software packages allow the operator to apply adjustments to the whole image or to certain parts of it. When the whole image is adjusted, it is clear to the viewer whether a draconian process has been used. For example, if a white glove is made to look green by application of a global (whole image) adjustment of color balance, the whole image will reflect the green tinge. On the other hand, if the glove alone is selected and then made to appear green, the rest of the image does not show any trace of the application of the color change. Clearly making local, as opposed to global, changes will require some good explanation to those viewing the image in order to assure the viewers that they are not being misled.

Some tools, such as the dodge and burn tools, are inherently local. They increase or decrease the brightness of a selected portion of an image. In fact, the edges of the selected area can be "feathered" so as to make the changes blend in with nearby, unchanged portions of the image. Unless there is a good reason for such a change it can be viewed as an attempt at deception. If the area to which the tool is applied is very small, on the order of a few pixels, then the use of the tool essentially is writing new pixels and should not be done. The cloning tool is even more of a problem. It copies pixels from a selected set of areas of an image and pastes those values into other selected portions of the same or different image. This is the tool that lets the operator put the head of one person onto the body of another. It is almost always an attempt at deception and it is hard to think of a valid reason for its use in forensic applications.

Another way to change portions of an image without changing the whole thing is to use one of the area selection tools. One such tool selects all the contiguous pixels within a certain value range; another selects all the pixels within a selected geometric shape. Once selected, tools that are otherwise global can be applied to the specific location. As with the inherently local tools, it is important that the operator—when making local as opposed to global adjustments—is prepared to tell viewers that this was done and why.

When image-editing software tools are used to change an image, mathematical functions are applied to the various portions of the original image and the result is shown as a new, modified image. The difference between the original image and the modified image can be thought of as a transparency. Conceptually it is like a layer that is placed on top of the original image. Image-editing software packages allow the operator to create layers, each with specific information to be added on top of the original image. Some are inherently opaque, like the addition of text or lines, and others are semitransparent, like density gradients. The value of having the modifications in layers is that the operator can change them easily to achieve the desired result without any changes to the original underlying image. When working in layers, it is best to keep the working image in that format. When the job is complete, the layers can be "flattened," or condensed down to a single layer that contains all the effects of the layering. The image formats of images with operational layers are proprietary and therefore not suitable for long-term storage, and they may not be compatible with other software packages such as those used for presentations. Thus when image adjustments are completed, the image should be flattened and saved as one of the more common standards, such as a TIFF.

SUMMARY

In digital imaging, every image is made up of pixels and each and every pixel is represented by a small set of numbers. These give the color for that pixel and its location. The location also provides proximity information – that is, the values for the neighboring pixels. Given this, it is possible to apply a wide range of mathematical functions to enhance the image, extract information, and compare images. Some of the processing techniques are mathematical representations of darkroom techniques originally developed for silver halide technology (for example: brightness, contrast, dodge and burn, unsharp mask and sharpen), some are taken from other sciences (for example, Fourier Transformation and histogram equalization), and still others are developed strictly for digital photography (for example: cloning and color space transformation. Although this was known before the days of digital imaging, it was only used on selected spots, not a full image). Most of these techniques operate on the full image, treating each an every pixel to the same adjustment processes. These are referred to as innocent adjustments. A few (for example: cloning, pencil writing) are specifically designed to make an image into something that it did not originally portray. These are called malevolent adjustments and should be avoided in forensic work.

The big advantage of digital imaging is the ability to employ all of these exotic image processing tools in seconds – and in some cases to use them at all. However, the user in a forensic setting must be careful to use the tools with full knowledge of how each alters the image. There should be good explanations of why certain tools were used and why others were not. Also, it is important to know the possible impact of the use of these tools beyond the intented purpose. This applies to the sequence of the application of tools as well.

EXERCISES

Rectification

Occasionally there will be images that were photographed at an angle. When this happens, it's almost impossible to size them accurately to 1:1 for examination. See Figure 10.E1.

FIGURE 10.E1

If a scale or other device has been inserted into the photograph, then you may be able to use the tools in PhotoShop to correct the perspective. Download image Original_5989. Note the horizontal and vertical distortion, caused by photographing it from an angle in the example.

1. Select the Measure tool from the toolbox.
2. Place a measurement line perfectly parallel with the bottom of the box as shown in the example in Figure 10.E2.

FIGURE 10.E2

3 Ensure that it is perfectly aligned with the bottom line of the box. If not, the rest of your adjustments will be off.

4 Select **Image Menu** > **Rotate Canvas** > **Arbitrary** from the menus.

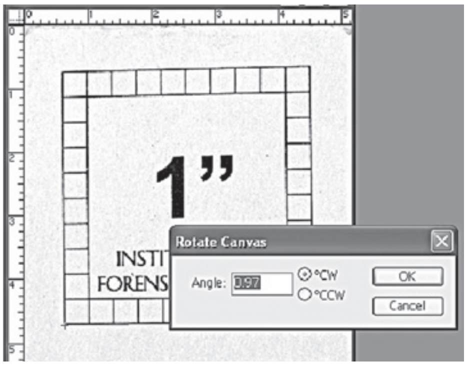


FIGURE 10.E3

5 The Rotate Canvas dialog box will appear, indicating the degrees of rotation needed to make the box parallel with the window as in Figure 10.E3. Click **OK**.

6 You must have the image parallel with the window prior to continuing. Double-check with the measure tool to ensure it is exactly straight as seen in Figure 10.E4.

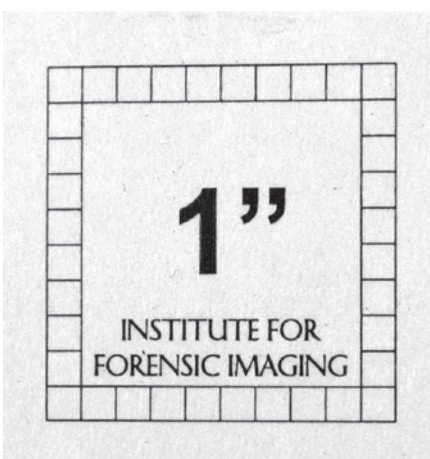


FIGURE 10.E4

Note: This is the most important step in the process.

7 Place two horizontal and two vertical guides on the image aligning them with the outermost lines of the box—at least two of them should align perfectly as shown in the example in Figure 10.E5.

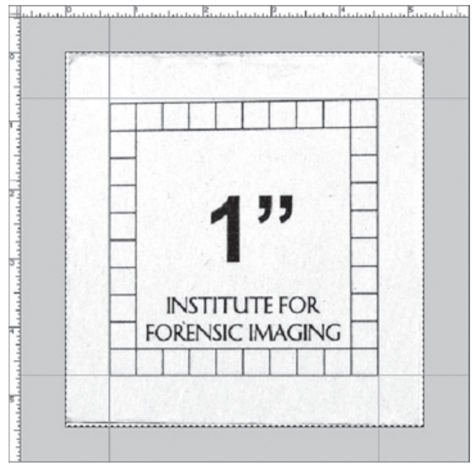


FIGURE 10.E5

8 Note the distortion in the upper left corner of the image. Enlarge the window by dragging the corner, placing some gray work area around the image.

9 Choose **Select** > **Select All** from the menus.

10 Select **Edit Menu** > **Transform** > **Distort**. Eight handles will appear around the image. The handles (usually the corners) are used to straighten the box to align it with the guides.

11 Drag the corner handle to align the box with the guide.

FIGURE 10.E6

12 Check each of the horizontal and vertical lines to ensure that they align with the guides as shown in the example in Figure 10.6. Press **Enter** to apply the changes.

Now let's look at a real image with a scale, as seen in Figure 10.E7. Download image, Cup on the website.

Note the picture in the example. Due to where the fingerprint was located, it was necessary to photograph it at an angle. Using the measure tool, place a straight line along the longest part of the scale as in Figure 10.E8.

FIGURE 10.E7

FIGURE 10.E8

There must be at least one parallel plane prior to beginning. The image will rotate similar to the example in Figure 10.E9.

1 Use a blue guide to ensure you have a straight line.

2 Crop the image to just the fingerprint and the scale.

FIGURE 10.E9

3 Place the guides in the image as shown in the example in Figure 10.E10.

4 Enlarge the window by dragging the corner, placing some gray work area around the image.

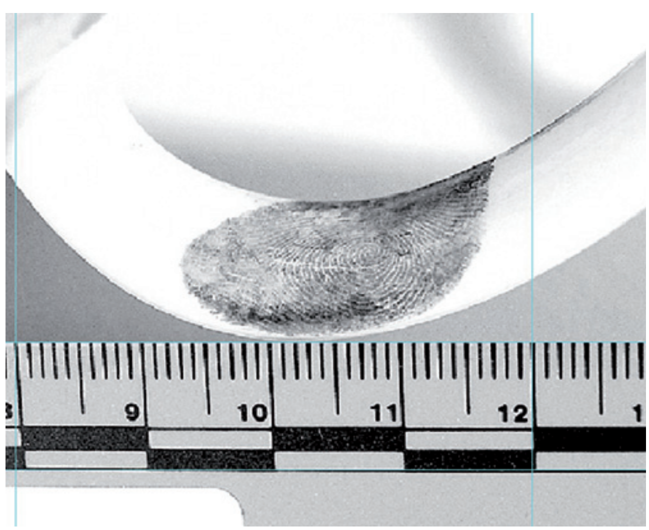


FIGURE 10.E10

5 Choose **Select** > **Select All** from the menus.

6 Select **Edit Menu** > **Transform** > **Distort**. Eight handles will appear around the image.

7 Use the handles to straighten the scale so it aligns vertically and horizontally with the guides as shown in Figure 10.E11.

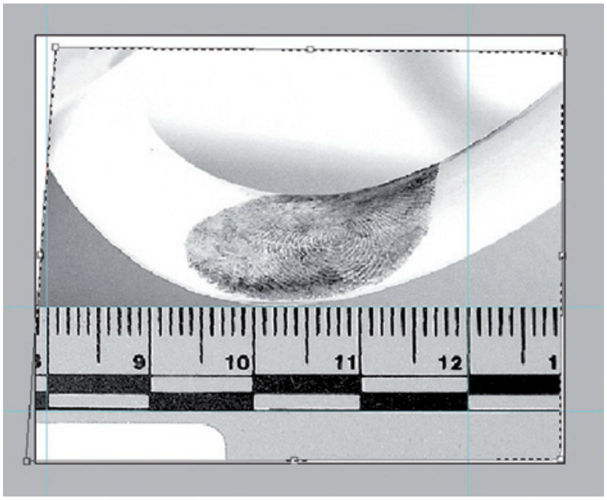


FIGURE 10.E11

8 Press **Enter** when finished to apply the changes to the perspective.

The image is now ready to be sized 1:1 for comparison, as in Figure 10.E12.

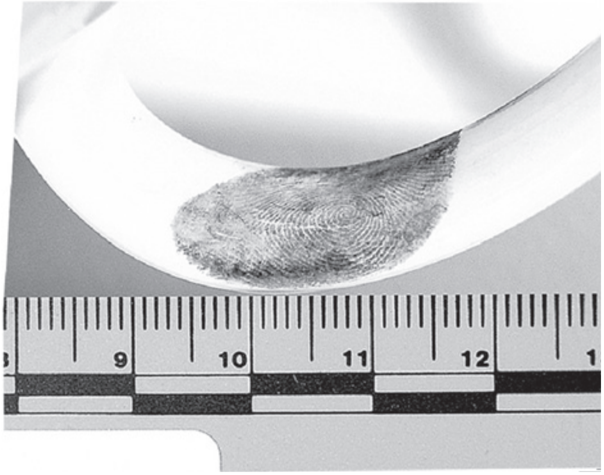


FIGURE 10.E12

Sizing 1:1

The Institute for Forensic Imaging (IFI) in Indianapolis, Indiana (www.ifi-indy.org) offers a sizing filter that can be installed in the plug-ins folder of Photoshop.

1 Open the image and select **Filters** > **IFI** > **Sizing**. The sizing dialog box will open, as shown in Figure 10.E13.
2 The first requirement is to establish a known unit of measure and a value. Refer to Figure 10.E14. The scale is indicating centimeters, so set the

FIGURE 10.E13

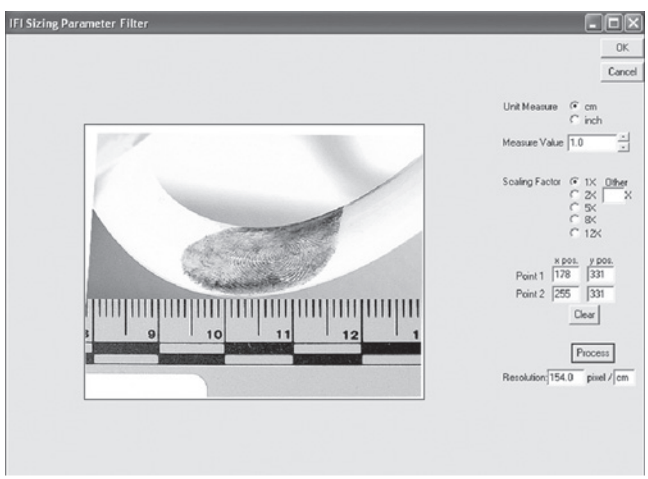


FIGURE 10.E14

unit of measure to "cm" and the value to 1.0. It can be scaled, but we're just setting it 1:1, so leave the factor to 1X.
3 Click the **clear** button to clear the x and y positions.
4 The next step requires two clicks of the mouse, one at the beginning of a centimeter and one at the end. If you accidentally click more than twice, clear it and start over. Note the value in the resolution box—your measurement should be approximately the same. If it is not, try again. Clicking OK does not size the image, you must do that manually. Note the value in the resolution box. Click **OK**.

5 Select **Image Menu** > **Image size**, as shown in Figure 10.E15.

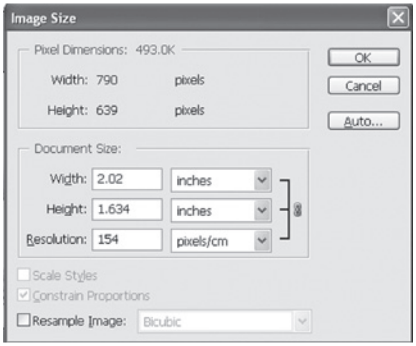


FIGURE 10.E15

6 Turn off the Resample.
7 Set the Resolution value to **pixels/cm**.
8 Enter the value **154** in the resolution box. Click **OK**. The image will be resized.
9 Immediately check with the measure tool to ensure that the image is 1:1, as shown in Figure 10.E16.

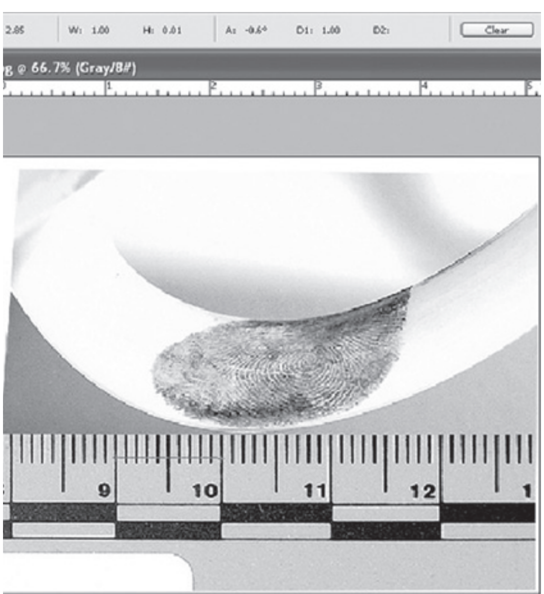


FIGURE 10.E16

10 Select the measure tool, set the unit of measure to **cm** using the info palette.

11 Use the measure tool to ensure that it is **one centimeter** in the options bar.

Making Measurements

Heights and distances can sometimes be calculated with this tool if you have a known value in the photograph and the perspective does not distort distances excessively. For complex settings reverse projection photogrammetry is a better tool. Download image SIZE from the website.

The known value must be very close and on the same plane as the object you are intending to measure. Study the image carefully and determine what measurement is needed to accurately make a measurement in the image.

FIGURE 10.E17

The counter in the sample picture, shown in Figure 10.E17 was chosen for the fixed measurement. Note the black line—this was inserted to account for the angle of view. The measurement made at the bank was 56″ from the top of the counter to the floor.

Once the known value is determined the sizing tool can be used, as shown in Figure 10.E18.

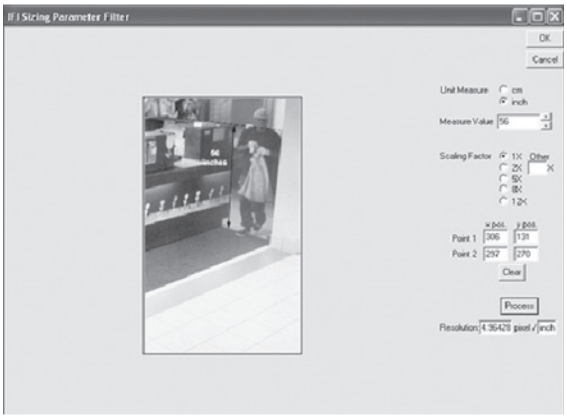


FIGURE 10.E18

1 Enter the unit and value as indicated in the example.
2 Indicate 56″ from the counter top to the floor as shown in the example.
3 Process the information as before and register the number for future use. Click **OK**.
4 Enter the information in the image size dialog box to set the image to 1:1.
5 Check the 1:1 measurement using the measure tool on the counter top. Refer to Figure 10.E19.

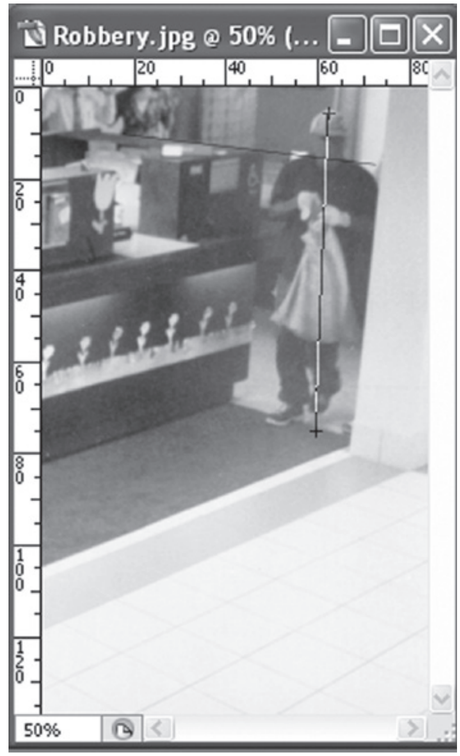


FIGURE 10.E19

6 To get the height of the suspect, measure from the top of his hat to the floor as indicated in the sample. Keep in mind there are a lot of variables, such as bent knees and wasted area in his hat.

CHAPTER 11

Scanners

When most people think about digital imaging, they think about digital cameras. Next are video cameras with digital outputs. Only a few think of scanners, even though many people use them. In the graphics industry—which deals with the production of magazines, advertisement flyers, and any number of materials—scanners are a mainstay. There are three main classes of scanners: drum scanners, flatbed (or desktop) scanners, and film scanners. Scanners record an image one line at a time by either moving the image relative to a stationary light source and sensor, or by moving the light source and sensor relative to a fixed image. There are also specialty devices such as scanner backs for cameras and laser scanners for three-dimensional photography.

DRUM SCANNERS

Drum scanners have been around for many decades and are the workhorses of the high-end graphics industry. They are capable of working from both reflection and transmission originals. These devices have a precision drum, shown in Figure 11.1, which rotates at high speed. The original is affixed to the drum. With reflection originals, a collimated beam of light is impinged on the top surface and a high-speed light-sensing device is used to capture the reflected light. For transmission work, the light beam comes from inside the drum and goes through the sample. Usually that sensor is a photomultiplier tube, since these have very high sensitivity and very fast response. They are capable of a wide range of light levels and a high-frequency response. The signal from the photomultiplier is sampled and digitized. These signals are synced to the angle of the drum when they were made in order to produce pixel information. To get the red, green, and blue records in a single pass, the device will have three photomultiplier tubes. Drum scanners can capture image data at some 12,000 pixels per inch and they can work with fairly large originals. For example, films that are 4 by 5 inches or larger will

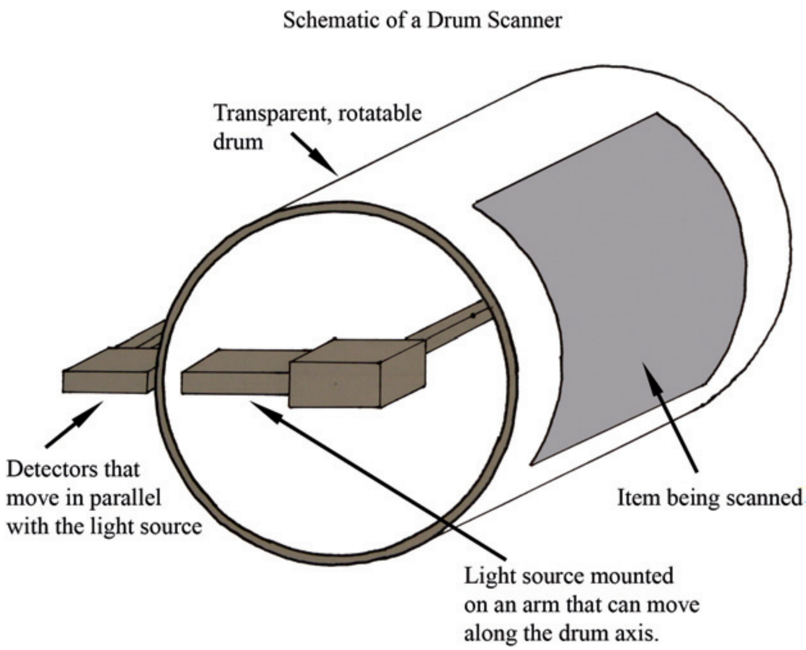


FIGURE 11.1 *Drum Scanner. A highly simplified schematic of a drum scanner is shown with the light source mounted on an arm that can move back and forth down the inside of the transparent drum. Moving along with the light source is the read head, which measures the light coming through the item being scanned as it rotates around with the drum. It is also possible to have a light source comounted with eh sensor to read reflection materials.*

probably have to be scanned on a drum scanner. Likewise, artworks that are a meter across will need to be scanned on such a device in order to get high-resolution information from a large original. Drum scanners can have separate controls to set the size of the area being scanned and the appropriate pixel dimensions. They can produce image files of several hundred megabytes; special computers are needed to work with such large files. These scanners are large and expensive devices and usually are used only for applications that require very high image quality. These applications include museum art replication, high-end magazines, and certain scientific reproductions. Until the advent of relatively high-performance film and flatbed scanners, they were the only scanner used in the graphics industry.

FILM SCANNERS

Scanners able to scan films in a flat mode started to become a force on the market around 1992, when the Eastman Kodak Company introduced a desktop unit that was based on a design for their PhotoCD system. They are capable of making excellent scans but usually are restricted to smaller formats such as 35 mm or 120 mm film. Larger originals will need to be scanned on a drum device. Typically film scanners have a resolution limit of about 5000 pixels per inch, or 197 pixels per millimeter. It takes three pixels to resolve a line pair, so this will correspond to 66 line pairs per millimeter. To put this into perspective, high-resolution films, such as Kodak T-Max 100 black-and-white film, are rated at 200 line pairs per millimeter under ideal conditions. This is accomplished under very special conditions and is not representative of what we might achieve in a high-quality 35 mm camera. Under the practical conditions, we are more apt to find a capability of 40 to 45 line pairs per millimeter. The result is that in a practical setting, the 5000 ppi scanner will be more than adequate to the task of scanning camera negatives. Many film scanners have a row of infrared sensitive pixels. The dyes used in photographic films do not absorb infrared light, so there should be no apparent image information in the infrared record. If any does show up, it is either a scratch in the film or dirt on the surface. This information can be used to correct for scratches and dirt.

SPECULA AND DIFFUSE ILLUMINATION

Scanning processes involve both illumination of the object to be scanned and collection of the resulting light. Light that is in an organized beam is called *specula*. It means that all the photons are moving either in parallel or possibly in a converging beam. Diffuse light is not organized. The photons are moving in many random directions. With collimated (or specula) light, shadows are preserved, whereas with diffuse light they are not preserved as much. Shadows are intense on a bright sunny day with its specula light and much less so on an overcast, cloudy day with its diffuse illumination.

Slide (film) projectors have specula light coming to the slide from the lamp house; the lens on the front that projects the image also creates a specula beam. This is a *specula–specula device*. They are designed to make a crisp image of the slide. A device that is designed to measure the amount of dye formed on a film after exposure and processing is called a *densitometer*. When used with a transparent sample, it is a specula–specula device. When used with reflection images, it is different. The incoming light is from a ring, and that light is brought in at 45 degrees. So it is diffuse around the point of study, but specula through the height of the hemisphere. The light collected is specula (a lens is used). This is called a *45–90* densitometer, which refers to the angles on incidence and collection. It is designed to minimize front surface reflections off of very flat surfaces, such as glossy photographic prints. When measuring the darkness or brightness of a nonglossy surface, it is sometimes better to use a device with *diffuse–specula* lighting. For example, a sheet of paper that has gone through a laser printer is illuminated by a diffuse source that emulates room light. These prints are viewed in normal room light. The collection is through a lens—that is, specula. This emulates the lens in your eye viewing the page. This is because normal office paper has a surface that will take a specula input light beam and render it primarily as a diffuse beam coming off the surface. When this is done completely, the surface is called a *Lambertian* surface. The toners that are used in laser copiers do not create a smooth surface as photographic paper does. Instead, these too are Lambertian. Scanners are very similar to densitometers and slide projectors.

When scanning film transparencies (negatives and slides), the mode of illumination can be important. The silver grains in traditional black-and-white films are able to provide very high resolution, but they also impart some degree of scatter to the light traversing the film. If the incoming light is an organized beam, the main effect will be the blocking of incoming light and this system will be capable of registering high densities at high resolution. If the incoming light is diffuse, the propensity for scattering is exacerbated. This will result in lower maximum densities and a bit of a loss of resolution. So for transparent materials, the best scanning will be accomplished with specula–specula geometry. If the sample to be scanned is a reflection print, then either 45–90 geometry, or diffuse–specula geometry, would be best. Drum scanners, which have very high performance requirements, are frequently used to scan photographic transparencies and so they use specula–specula geometry. Most flatbed scanners are designed primarily to scan sheets of normal paper. These typically have diffuse–specula illumination.

Film scanners use solid state sensors. These are charge-coupled devices in which the pixels are all in a single-file line. Typically there are four such linear arrays on the chip or chip set: one each for red, green, blue, and infrared (refer to Figure 11.2). There are filters affixed to the pixels. The sensor chip and the lens are on a single mount and move on a precision track while the sample is held still. Most scanners can work with both film strips and mounted slides. An important aspect of dedicated film scanners is that the light that illuminates the sample is collimated. This means that any scatter imparted to the light will add to the density ascribed to that portion of the film.

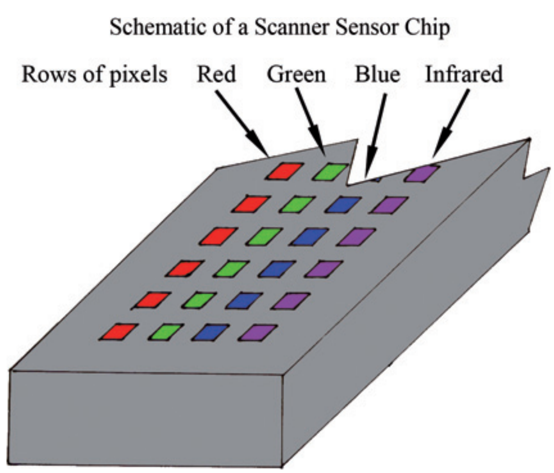


FIGURE 11.2 *Scanner Sensor Chip. These chips are called bar or linear arrays. The pixels are arrayed in lines and receive their respective lighting as the device moves relative to the original. It captures a row at a time.*

As with drum scanners the analog readings taken from the sensor chip are converted to digital and linked up with information that identifies the location of the portion of the film for each reading. The files from film scanners can be fairly large. For example, a scanner with 5000 pixels per inch scanning a 35 mm negative will create approximately 26.3 megapixels. Each pixel has three values and these are two bytes deep (typically). The result is a file of 158 megabytes, assuming no compression. Film scanners are in wide use, but the need for new ones has dropped off as the sale of film has decreased.

FLATBED SCANNERS

By far the scanners most in use are flatbed, or desktop scanners. They come in dedicated scanner form or they might be built into a fax machine, a copier, or some other multifunction device. This widespread penetration of the market has resulted in very low unit costs. These devices are often capable of scanning both reflection materials and transparent originals. The devices hold the original still on a glass platen and have a lens-and-sensor housing that moves along a precision tack (shown in Figure 11.3). As the housing is moved along the track, it takes a snapshot of a line of the image. The housing moves in discrete steps and the lens and sensor design is such that a row of pixels is measured at each step. The information is held in a buffer until the scan is complete and then it is compiled to create an image.

The limitations on resolution depend on the direction of travel. Movement in the direction of the tracks has resolution limited by the size of each step (in combination with the lens and pixel size); resolution in the

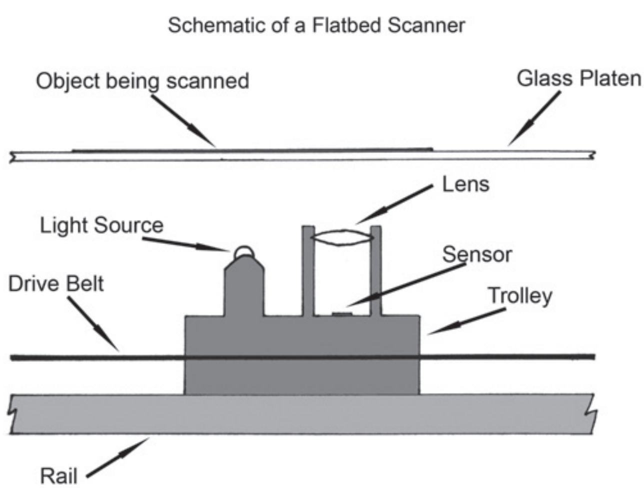


FIGURE 11.3 *Flatbed Scanner: At the top, an original is laid on a glass platen. Below this is a trolley that carries a light source, a lens, and a sensor chip (as in Figure 11.2). The trolley is pulled along by a drive belt and rides on a precision rail system.*

dimension across the scanner is limited only by the pixel size and spacing in combination with the lens. These limits are referred to as the *optical resolution* of the scanner. But scanners, like all digital image-capture devices, apply some image processing in order to deliver a viewable image. It is altogether possible that the software used to create the delivered image is able to interpolate the image to either lower or higher resolutions. If the device is rated as being able to deliver images with greater resolution than the optical resolution, then that image has been interpolated up. In other words, a certain number of pixel values were estimated from near neighbors, and they are not a direct result of the content of the object being scanned. It is generally advisable to seek a scanner with an optical resolution equal to the highest resolution the laboratory will want to routinely capture.

SCANNERS vs. DIGITAL CAMERAS

In general, scanners give better resolution, dynamic range, and color fidelity than digital cameras. At the same time, they have a limited object size and shape, and they require an object that is not moving. Scanners with optical resolutions of 1000 to 1500 pixels per inch are readily available for less than $100. For a slight increase in price, you can purchase a scanner with an optical resolution approaching 5000 pixels per inch. Since it takes three pixels to render a line pair, this translates to 1667 line pairs per inch, or 66 line pairs per millimeter. This is greater than the practical limit on commercially available high-performance film and camera combinations. If the platen is 8.5 by 14 inches, the scanner can capture an image that is 42,500 by 70,000 pixels. This translates into a camera with 2975 megapixels—far more than any commercially available digital camera is likely to have, now or in the future. So for relatively flat objects that are less than 8.5 by 14 inches and for which a very high resolution is required, a scanner is the best tool.

Scanners have a built-in light source and fixed optical alignment. This results in an intrinsic ability to give consistent color rendition and higher dynamic range. The color consistency is a result of the fact that the light source is always the same and the system's color response can be easily calibrated. Also, since most scanners use three separate sensor arrays for the visible colors, there is no need to estimate the image, as is done in most digital cameras (the demosaic process described in Chapter 16). There is no need for white balancing, for example. Since the light path is always the same, it can be designed to minimize front surface reflections (per 45–90 densitometers). The result is that higher density readings are possible.

There are three main drawbacks for scanners. First of all, the object must be flat or nearly so. If it deviates by more than a centimeter from the platen, there will be a tendency for the image to become blurred, dark, and distorted. So even though you can use a digital camera to capture an image of a doorknob, it is much harder, if not impossible, to do so with a scanner. The second drawback is size. You can take a photo with a camera of a section of a wall with blood spatter that is 4 feet by 4 feet. This will not be possible with a scanner. Finally, the object must be still during the scan. A very high-resolution scan can take several seconds, and if the object moves during this time, the image will be distorted and blurred.

SAVING OUTPUTTED IMAGES

When a digital camera takes a photo, the image file is sent to a removable medium so that the image can be taken into a computer or printer. That file contains much more than just the image. It also has a complement of *metadata*, including which camera was used, what all of its settings were, and when the picture was taken. If the camera has a GPS attachment, it will tell where the photo was taken. Scanners only send out the image. The scanner is attached to a computer and the computer is running software specifically for that scanner. The raw data from the scanner is converted to an image in the computer and a file is created. Unlike cameras, no metadata is attached. It is up to the operator to ensure that important image provenance information is recorded so that it will be possible to associate that data with each specific image.

EXERCISE

Scanner Types

35 mm Film scanners

35 mm film scanners range from around $250 and up. The more expensive scanners:

- Usually have higher resolution
- Have a higher D-max
- Many now have infrared dust removal

Multipurpose Flatbed Scanners for 35 mm, Medium Format, and 4 × 5

Most of these can scan reflected images up to 8.5 inches wide, although you can purchase more expensive models that scan larger images. They usually come with transparency adapters that are capable of scanning various sizes. The price varies from $200 and up. They are not as good as dedicated film scanners, but they are adequate for 35mm enlargements up to about 8 by 10 inches. The larger the original negative, the larger the print you can make.

Dedicated Medium Format Film Scanners

These scanners have resolution of 4000 dpi and an excellent Dmax. Most of them have a Dmax = 4.8 or greater.

Purchasing a Scanner

There are three important areas to watch for when purchasing scanners:

- Resolution in DPI (optical and enhanced)
- Density range (Dmax)
- Software

Resolution

There are two types of resolution available with all scanners:

- Optical—the only true resolution obtainable by the scanner.
- Enhanced—requires interpolation by software provided with the scanner.

For any resolution above the optical rating, the hardware/software of the scanner will have to interpolate the image to add pixels, thereby affecting the quality of the final product.

Density Range (Dmax)

The capability of the scanner to capture the full density range of the original image. Maximum Dmax on any scanner is 5.0.

- Film Scanners—almost always have a higher Density Range than a flatbed scanner. Maximum Dmax range on most film scanners is from about 4.2 to 4.8.
- Flatbed scanners—usually have a maximum Dmax of about 4.0.

Software

There are two types of software provided with scanners:

- Amateur—allows limited control over the scan by the operator. Cropping and resolution are usually adjustable; exposure, color balance, and such are usually all in auto mode.
- Professional—allows full control over the scan. The operator will have most of the same controls that are available on a digital camera.

A good point to remember when scanning is "Garbage In, Garbage Out." Strive for the best scan possible, just as if you were taking an original photograph in the field.

Scanners and Bit Depth

Most standard image file formats store pixels with 8-bit precision. This is more than sufficient for full-toned prints with gradations as fine as the eye can see, but it is well below the capability of most scanners. Tonal detail is lost in converting a 10-bit (and higher) scan to an 8-bit file. If the pixels in the scan correspond closely to the desired tones in the print and if little editing is required, the loss is not significant.

But if significant editing is required, there can be a loss of tonal gradation and subtlety. This degradation can be avoided by using file formats with 16-bit precision (48 bits per pixel for color). No information is lost—images can be edited extensively without loss of tonal quality.

Most scanner software, but not all image-editing software, supports 16-bit images. Photoshop offers good support for 16-bit images.

Scanning Objectives

The high-end scanner software gives you considerable control over the appearance of the scanned image, including:

- Resolution
- Cropping
- Color balance
- Saturation
- Brightness
- Contrast
- Sharpness

The adjustments are made prior to the final scan. The resolution, exposure, and cropping are applied to the scan itself. The remaining adjustments are applied after the scan but before saving the image.

Color and tone adjustments are done with a slightly different intent, depending on whether you are saving the image with 8- or 16-bit precision.

- With 8-bit scans, significant editing in an editing program causes loss of detail. Therefore the scan should closely resemble the intended final image.
- With 16-bit scans, you should maintain detail with the scanner in the shadows and highlights. Typically, little adjustment is required; automatic settings are designed to accomplish this. It doesn't matter if tones don't precisely match your intent; you can easily adjust them in the image editor without loss of detail.
- 16-bit scans are much more forgiving. This is strongly recommended for maximum quality.

Resolution depends on:

- The scanner's optical resolution
- The quality level you required at the maximum intended magnification

Each scanner has a maximum optical resolution, specified in ppi, often called dpi. Many scanners often have interpolated resolution numbers that are much higher. Do not use these interpolated resolutions. They only make larger file sizes and offer no advantage.

If you need to resize, do it later in your image-editing program.

The formula for determining the scan resolution is as follows:

Compute the magnification first:

Magnification = M = print dimensions/ film dimensions

For example:

- The size of a 35 mm negative is (24 mm × 36 mm, which is equivalent to 0.945 inches × 1.42 inches).
- The size of the paper is 8 ½ × 11 inches (10.5 inch image length)

Calculate the magnification:

$$M = 10.5/1.42 = 7.3x.$$

If you crop the image you will need to recompute the magnification. Larger paper sizes will also increase the magnification.

To determine ppi required for the magnification:

Scanner ppi = M * print ppi

The majority of printers will print at 300 ppi for the highest-quality print, but this can vary with each printer. Always check the specifications on your printer.

Continuing from the previous example:

Calculate ppi = 7.3 * 300 = 2190 ppi

Remember that images scanned at high resolution intended to make sizeable, high-quality prints can be quite large, storage-wise. A full-frame 35 mm image scanned at 4000 dpi in 48-bit color requires about 128 megabytes.

The chart in Figure 11.E1 indicates the results obtained at different resolutions. But always run tests on your printer.

Pixels Per Inch sent to printer	
	Description
100	Adequate for very large images viewed at a distance.
150	Good for big enlargements-- 13x19 inches or larger.
200	Excellent sharpness for all sizes; nearly as good as 300 PPI.
300	As sharp as the eye can see; no advantage to increasing PPI.

FIGURE 11.E1

Scanner Software

Scanners operate under the control of software, which is supplied by the manufacturer. Scanner software can be divided into two categories: professional and amateur. The amateur software is not acceptable for high-end scanning, because it usually does not allow control over enhancements of the image. It lets you set the resolution and do some cropping, but then it is a one-button automatic exposure scan.

Most professional software packages allow you to scan into an image-editing program, such as PhotoShop, using a TWAIN interface or by going directly into the software program.

Here are some questions to ask when comparing scanner software packages:

- Does it have a histogram? Histograms are very useful for making tonal adjustments to ensure that the shadow and highlights are set correctly.
- Does it support 48-bit scanning?
- Does it have controls for color balance and sharpening?
- Does it allow you to save your exposure settings so that they can be applied to additional scans?

Scanning Steps

Scanners and software packages differ in details, but the typical steps are similar.

1 *Clean the film*. This is a critical step. Clean the film carefully, examining it closely under a bright lamp. Canned air can be used and, if necessary, a soft camel-hair brush.
2 *Check for infrared dust*. Some scanners may have infrared dust (IR) removal, in which the presence of dust is sensed by an IR light beam, and the offending areas are filled by interpolation.
3 *Open the scanner software*. Most professional scanner software operates standalone. Some can be opened from within the image-editing program via the TWAIN interface.
4 *Insert the film into the scanner*. Most scanners have film holders, but some film is put directly into the scanner.
5 *Set the correct film type*. Color positive or negative; black-and-white negative.
6 *Set the resolution and size*. Most scanners have a control for the resolution and magnification.
7 *Scan the thumbnails*. This step scans the film at low resolution and displays small thumbnail images.
8 *Select the image(s) to be scanned*. Depending on the software and preferences, either select specific thumbnail(s) image(s) you want to scan or allow all to be scanned.
9 *Adjust the image(s)*. This is where you enhance the image(s) via: resolution, exposure, cropping, tone (brightness and contrast), and color (balance and saturation).
10 *Perform the final (full resolution) scan*. This can be slow for high-resolution scans. If an IR dust removal is employed, it may require two scanning passes.
11 *Save the image to a file*. Save high quality images in the TIFF (.tif) format.

Scanning Example

This example uses Hewlett-Packard's PhotoSmart S20 scanner, opened from within HP's software. It can also be accessed through PhotoShop using the TWAIN interface.

1 Access the software via the HP utility.
2 This brings up an empty Preview Scan screen. Set the photo type and scan resolution.
3 Clean the film and insert it into the scanner. The thumbnail images (in this case, for a strip of four) appear in the Preview Scan screen in Figure 11.E2.

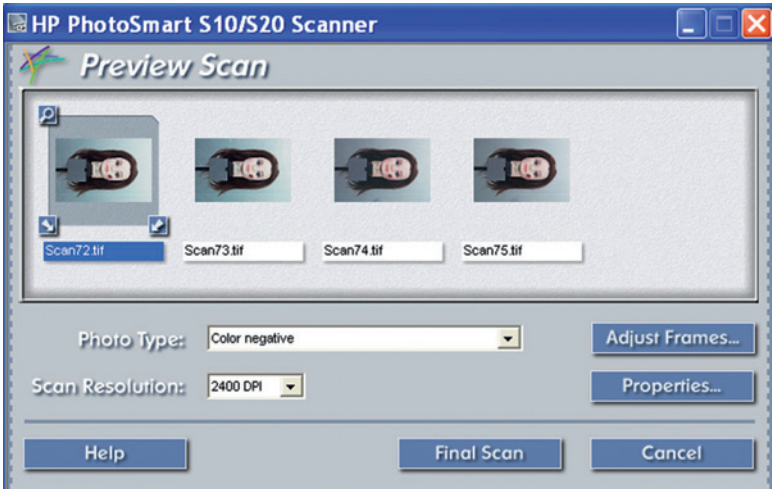


FIGURE 11.E2

4 Select the image to edit. Small icons appear around the thumbnail. The two on the bottom are for rotating the image. The one in the upper left accesses the Image Adjustments screen, a portion of which is shown in Figure 11.E3.

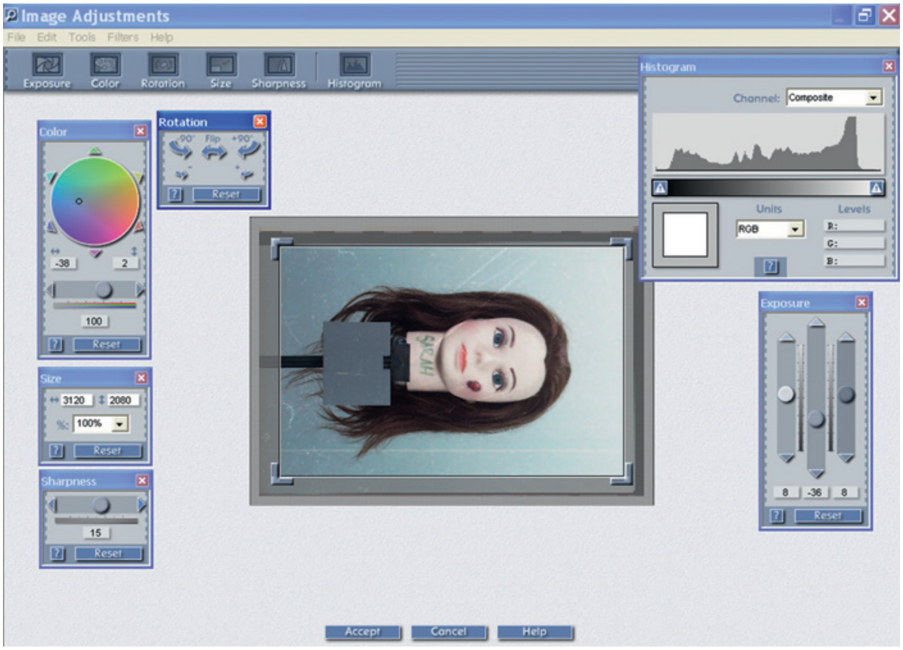


FIGURE 11.E3

This screen includes a Histogram with exposure control. It also includes controls for sharpness, rotation, size, and color.

The PhotoSmart software makes its best guess for color balance and saturation, which is set in the Color box. Brightness and contrast are set in the Exposure box in conjunction with the histogram.

Cropping: This is controlled by the four handles around the image. It is usually good practice to keep a little more of the image than you expect to use. Fine cropping can be done in image editing.

Color balance and saturation: This is controlled by adjusting the color wheel.

Exposure box settings [highlight, middle, shadow]: Using the Exposure box controls, the Histogram screen can be used to set the levels in the image. The Exposure box has controls for the highlights, shadows and midtones, as shown in Figures 11.E4 and 11.E5.

Sharpness: The default value is 15. Sharpening is difficult to control at this stage—you can't see its actual effect because the preview scan has lower resolution than the final scan. As a rule, it is best to leave the scanner at its default setting and the do the final sharpening later in the image-editing process with the Unsharp Mask.

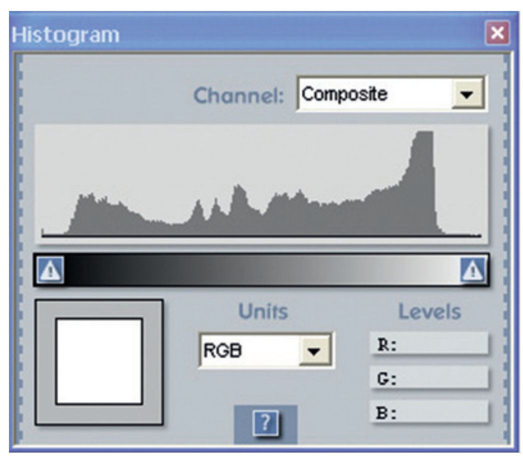


FIGURE 11.E4

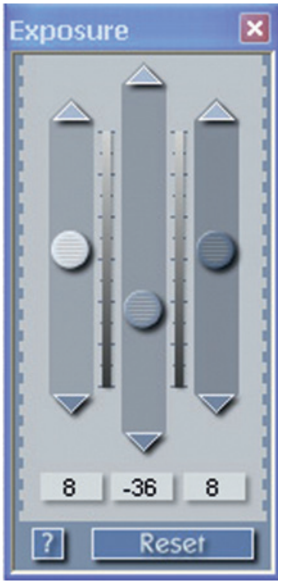


FIGURE 11.E5

When you are satisfied with your settings, click **Accept**. This brings back the Preview Scan screen. Make sure all the images you want to scan are selected and you have the correct scan resolution, then click on **Final Scan**. A 2400 dpi scan is rather slow. Quality, not speed, is the objective.

Digital Circuits

Digital photography offers the twin advantages of not requiring the consumption of supplies for every photo taken, and the ability to show the image within seconds. An additional benefit has been the ability to easily and quickly enhance images, extract information from them, and reliably move them from one location to another. In order to do this, many complex electrical elements are employed and complicated mathematics is used to assure high-quality results. In this chapter the more basic electronic circuits will be described. And since these are predicated on binary arithmetic, this too will be explained.

BASIC ELECTRICITY

Electricity is the movement of electronic charge. Electrons have intrinsic electric charge, and when they move along a path, they comprise an electrical current. The impetus to get the charge to move is electrical potential. Current is measured in amperes (or amps) and electrical potential is measured in volts. By way of example, as a big rain storm builds, rain drops form, and when they start to fall the process causes an electrical charge to accumulate in the clouds. Since there is no such activity on the ground, a voltage will grow between the clouds and the ground. When the voltage gets so high that the intervening air breaks down, the charge in the clouds will flow to the ground in a lightning bolt. The voltage existed before the current flowed since voltage, as always, is measured between two points. The flow of the charge between the clouds to the ground constitutes an electrical current in the form of lightning. Each bolt dissipates some of the accumulated charge.

So current is the flow of electrical charge through something and voltage is the electrical force between two points that causes the current to flow. The product of the two is the electrical power, measured in watts—watts equals amps times volts. High voltage and low current results in low power. An example is walking across the carpet in a dry room and then touching the light switch. The voltage is quite high, but the current is very small and so no real damage is done. Another example is connecting a household light bulb across a car battery. The bulb is designed to work at 110 volts, and the car battery produces only 12. Current will flow, but the bulb will not glow. The current is significant, but the voltage is low and so there is not enough power to make the bulb light up. Current goes through and voltage is across. Voltage never goes through!

There are two passive devices that are used in electrical circuits that are of interest in this discussion: the resistor and the capacitor. A resistor is simply a device that has limited ability to conduct current. This ability is built in by the nature of the materials used and the geometry of the element. Electrical resistance is measured in ohms, where one ohm is equal to one amp per one volt. That is, if one volt is put across a one ohm resistor, one amp of current will flow. If the voltage is increased to three volts, the current will go up to three ohms, and so on. The voltage is measured across the resistor and the current goes through the resistor. The power dissipated is the product of the amps times the volts. The more resistance an element has the less current it will conduct for any given voltage. It is important to note that high resistance (or sometimes a related term is used, impedance) prevents high currents, and therefore only slightly dissipates any accumulated charge. High resistance also often is associated with low power consumption. This is why power lines are operated at very high voltage.

A capacitor is like a jar that holds electrical charge. It has two large area plates separated by a membrane that does not conduct electricity (an insulator). Charge can accumulate on the plates in response to the imposition of a voltage. The charge flows to the plates, which causes a temporary current, but current does not flow across the insulator membrane. It just sits on the plates until it is discharged by some external means. Electrical charge is measured in coulombs, capacitors are measured in farads, and one farad is equal to one coulomb per volt. So, if you know the size of a capacitor (in farads) and measure the voltage across its terminals (with a very high resistance device so that no appreciable current flows), you can infer the amount of accumulated charge (in coulombs).

Photo detectors such as those used in the sensor chips in digital cameras are like small capacitors, each connected to a charge producing device. The charge producing device in this case is driven by the incoming light. The number of coulombs generated is proportional to the number of photons incident. And since the capacitance (in farads) is known, we can use the voltage as a direct indicator of the incident light. All of this is an analog process. The voltage (a physical entity value with no number) from each and every photo detector, or pixel, will later be converted to a digital value (a numerical entity with no physical component).

To apply this, consider one of the differences between CCD and CMOS sensor chips. In charge coupled devices, electrical charge is made to flow along a sequence of physical locations by the imposition of a series of pulses of voltage.

This results in the consumption of a noticeable amount of power. In complimentary metal oxide on semiconductor sensor chips, the current from each pixel location is converted to a voltage and then to a number right at that location. The only thing that moves from that point on is a series of voltage pulses, each with very low current. As a result, very low power is consumed. Cameras with CMOS chips are much easier on camera batteries than those with CCD chips.

ENTERING THE BINARY WORLD

Once an image is numerically encoded—that is, the analog values from the sensor chips are converted to discrete numbers—it can be distributed, verified, enhanced, compressed, merged, and easily processed in many ways that are difficult to accomplish in analogs. One of the fundamental building blocks in the world of digital systems is the electronic bi-state circuit. The output of this basic element is a voltage that is either on or off. The on state is called 1 (one) and the off is called 0 (zero). There is nothing in between. The voltage variation makes a sharp jump between two allowable states; this is accomplished by the *pulse-code modulation* process (PCM). The PCM compares the input signal pulse against a certain threshold value, if the signal is greater than the threshold, it is assigned the high value—one; otherwise it is assigned the low value—zero. Sometimes, this process is referred to as *binarization*, since it separates a signal into a binary value of either one or zero. It is the answer to the most fundamental question—is this value greater than the reference value? And it is called a *bit*.

As an example, let's consider the examination for a driver's license. The actual score can vary between 0% and 100% correct. But the requirement is a score of 70% to get a license. There are only two ultimate consequences, pass or fail, once the exam is completed. We need only one bit to represent this test outcome. Since the department of motor vehicles set the threshold score as 70%, we compare the score to 70% and if it is at that level or more, the output bit is set to 1. This means the examinee passed the test. Any score less than 70% will result in the bit being set to 0 and the examinee has failed.

To represent additional classifications, for example both the written exam and the driving exam, more than two states are required. That is to say, one bit is not sufficient. In this simple example, two bits can be used. In general, any number of combinations of multiple bits can be employed. Four different states can be made from 2 bits since $2^2 = 4$. *Bit patterns* are made by displaying the values for the multiple bits in sequence. In this simple example where we are using two bits, the possible bit patterns are:

00: Failed both the written and the driving exams
01: Failed the written but passed the driving
10: Passed the written but failed the driving
11: Passed both the written and the driving exams

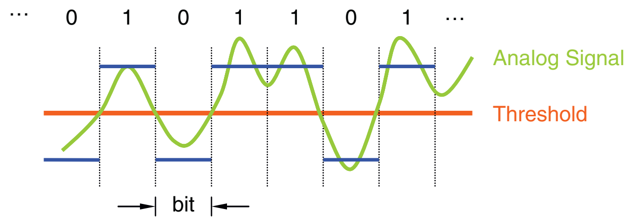


FIGURE 12.1 *Conversion from an analog signal into a bit pattern.*

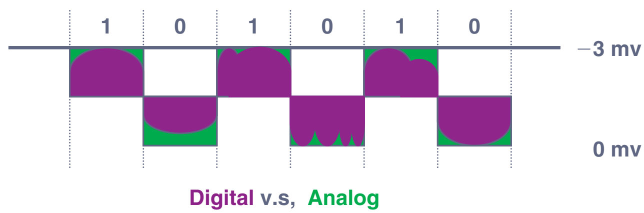


FIGURE 12.2 *Digital signal (square wave function in green) has less degradation than analog signal (sine wave function in blue). Using digital signal for communication can prevent signal drifts.*

Computer systems use 1 byte, the bit pattern with a length of 8 bits, as the basic information representation. An 8-bit binary string can represent $2^8 = 256$ different bit patterns from 0000 0000 = 0 to 1111 1111 = 255 (it is common practice to leave a space between groups of four binary digits each to make it easier to read the numbers). Each pixel in a normal grayscale image takes exactly 1 byte or 8 bits. Therefore, it can represent 256 different gray levels, starting from value 0 (black) to 255 (white). In general, n bits can represent 2^n binary bit patterns. Some camera images are represented in 16 bit pixels, but the principle is the same. Figure 12.1 illustrates how an analog signal is converted to a bit pattern.

Binary representation is generally more reliable than analog representation since it has only two values, 0 and 1, and it is easy to distinguish between these even when some noise is mixed into the signal. The result is less error when signal "drifts" are embedded with noise. Figure 12.2 illustrates the digital signal has less degradation than the analog signal, which has signal drifts.

DATA REPRESENTATION IN DIGITAL SYSTEMS

One important aspect of digital system design is how information ultimately will be converted into a bit pattern. Numeric values are the most prevalent and natural type of data representation from an analog source. When converting numeric information to binary codes, it is necessary to have a good mapping strategy, so that the data can be well represented relative to how it will be used. Numeric information can be categorized into three data types:

- Integer
- Negative value
- Floating point representation

This section discusses the representation of these three data types.

Integer Representation: Decimal versus Binary or Hexadecimal

To introduce binary data representation, it is useful to dissect the more common decimal system we all know. The decimal number 123 is based on powers of 10 (hence the term decimal, from the Latin for ten). It is really 1 times 10^2 plus 2 times 10^1 plus 3 times 10^0. or $1 * 100 + 2 * 10 + 3 * 1 = 100 + 20 + 3 = 123$. Each place holder is multiplied by 10 raised to a power consistent with its place in the string of digits, and the value in each place holder can be any whole number between zero and 9. The powers of 10 increase by one, going up from zero, as one progresses to the left. Similar rules apply in the binary system but the values are different. Each place holder is multiplied by 2 raised to some power, and the value in each place holder can be any whole number between zero and one. The powers of 2 increase by one, going up from zero, as one progresses to the left.

Binary is represented by only zeros and ones. A binary pattern (number) can be converted to a decimal number and a decimal number can be converted to binary. The binary number system is sometimes called the base-2 system, and base might be indicated by 2, b, B, or Bin. So 1011_2 or 1011_B are both binary numbers. The digit in each place is weighted by the appropriate power of 2.

To represent the value of 34 in the binary number system, we can apply the divide-by-2 technique. Referring to the pattern in Figure 12.3, 34 divided by 2 is 17 with no remainder; 17 divided by 2 is 8 with a remainder of 1; 8 divided by 2 is 4 with no remainder; and so on. Note that each place holder has to be a whole number (integer), and so the process ends when a division by 2 results in no whole number and a remainder.

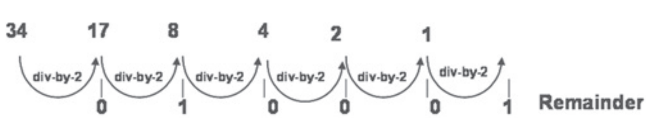


FIGURE 12.3 *Illustrate a common practice of how to use divide-by-2 strategy for converting a decimal number 34_{10} to a binary bit pattern 100010_B.*

Table 12.1 *Process of Binary to Decimal Conversion. A Binary Pattern 0101 Equals to Decimal Value 5*

Bit Location	Bit 4		Bit 3		Bit 2		Bit 1
Weight	$2^3 = 8$		$2^2 = 4$		$2^1 = 2$		$2^0 = 1$
Bit Pattem	0		1		0		1
	0	+	4	+	0	+	1

The next step is to collect the remainders from *right to left*. The equivalent binary number for value 34 is 100010_2. The mathematics behind Figure 12.3 is to decompose 34 into a series of 0s or 1s weighted by a power of 2 in increments from right to left position. The decimal value 34 can be rewritten as

$$34 \equiv 1\times 2^5 + 0\times 2^4 + 0\times 2^3 + 0\times 2^2 + 1\times 2^1 + 0\times 2^0 \equiv 100010_2$$

The conversion from the binary number system to the decimal system follows a similar process but the steps are reversed. It can be done easily if the weights can be applied to the corresponding positions appropriately where 1s are. For example, the process to convert 1011_2 to an equivalent decimal number is $(1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (1 \times 2^0)$, equals $8 + 0 + 2 + 1$ and the result is 11. The decimal number 11 in the decimal system is equal to 1011_2 in the binary number system.

For another example, convert a binary pattern 0101_B into its equivalent decimal value. Follow Table 12.1, which illustrates this process.

After summing the product of bit patterns and their corresponding weights, the decimal value of binary pattern 0101_B is 5. In general, the very right bit, bit 1, is called the *least significant bit* (LSB) because its weight is only 1, which has least contribution to the final value. The very left bit, the fourth bit in this case, is called the *most significant bit* (MSB), which has the most significant impact on the final value.

From Table 12.1, we notice that the values of weight at corresponding positions actually are universal. The weights from LSB to MSB, right to left, are always $2^0 = 1$, $2^1 = 2$, $2^2 = 4$, $2^3 = 8$, $2^4 = 16$, $2^5 = 32$, 64, 128, 256…it keeps going with the incremental powers of 2. Toward this end, we can utilize this property to expedite decimal-to-binary conversion. Let's try to convert the value 23 to its corresponding bit pattern. We first write down this universal weight sequence.

…	64	32	16	8	4	2	1

From left to right, we pick a weight value less than but closest to 23. The number is 16. We then mark 1 underneath weight 16, such as:

…	64	32	16	8	4	2	1
			1				

Now subtract 16 from 23 to get 7. And we now continue the same process and work on value 7. The next largest weight value that is less than 7 is 4. We then put another 1 underneath weight value 4. We continue on its remainder 3 after 4 is subtracted from 7. And weights 2 and 1 will be chosen if the same process keeps going. After it is done, we fill zeros into empty spaces.

…	64	32	16	8	4	2	1
			1	0	1	1	1

The final result of bit pattern for value 23 is 10111_2.

Since binary strings can sometimes be very long and contain many 0s and 1s, they can be very tedious to notate. For this purpose the hexadecimal number system is commonly used. It is used to represent the binary values for computer memory contents, address allocation, and digital image (color) pixel values. The hexadecimal number system is a base 16 system and uses 16 symbols 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A(=10), B(=11), C(=12), D(=13), E(=14), and F(=15) instead of 0 to 9 used in the decimal system, or the 0 to 1 in the binary system. Its bases can be written as base-16 or symbol H to be distinguished from decimal numbers, or plain binary numbers. For example if a pixel's gray level $2F_H$ this would be $2*16^1 + 15 * 16^0$, or 32 + 15, which is equivalent to the value of 47 in decimal. To convert a bit pattern to its corresponding hexadecimal number is very straightforward. We can group every four bits from LSB to MSB and fill zeros if there are less than four bits for the MSB group. For example, given a bit string 10 1001_2, it can be grouped by every four bits from right to left, such as

$$\underline{00\mathbf{10}}\ \underline{\mathbf{1111}}_{\ 2}$$

We apply binary-to-decimal conversion individually for each four-bit segment. Note that the A ~ F alphabetic characters need to be used for values greater than 10.

$$\frac{00\mathbf{10}}{2}\quad \frac{\mathbf{1111}}{F_{(=15)}}\ _2$$

In this representation, 0010_2 is equal to 2 decimal, and 1111_2 is equal to 15 decimal. The final hexadecimal code for binary 0010 1111_2 is $2F_H$.

The conversion between decimal and hexadecimal values involves multiple consecutive divided-by-16 processes and it can be a tedious calculation if the number is large. Therefore, the binary pattern often is used as intermediary in the conversion.

Representing Negative Values

In the decimal system, we simply put a minus sign in front of a number to indicate that it is a negative number. This is certainly a valid way for negative value representation. This approach is called *signed magnitude* and was used in most early digital or computer systems. It is quite straightforward and easy for circuit design—just add a circuit wire for sign indication. However, it can cause some confusion. Consider the following examples. Assume we reserve the MSB as a sign bit where 0 is positive and 1 is negative. Further assume that the length of bit string is 5 bits. So the bit representation can range from 0 0000 to 1 1111, and the corresponding value representation is shown in Table 12.2.

We can immediately see where the confusion is. There are two zeros. The representation is +0 (0 000) and −0 (1 000), and this does not make sense.

The modern binary representation is based on the 2's complement method. The sign bit is still at the same MSB as in the signed magnitude method; however, the remaining bits are complemented first and a 1 is added to the completed digits. The term of complement is the operation of flipping bit 0 to 1 or 1 to 0, which is similar to turning a light switch off to on or on to off. Let's use the −7 as an example to exercise 2's complement binary conversion. Assume the length of bit pattern is 8 bits, which is also called a *byte* in a digital system.

Figure 12.4 shows the process of how a negative integer −7 decimal is converted to a binary value. We first ignore the minus sign and only consider that the magnitude 7. 7_{10} is converted to the binary equivalence which is $0000\ 0111_2$. At this point, the MSB is always 0 because 7 is a positive number. The next step is to flip MSB signed bit from 0 to 1 and apply complement operation only on the remaining (magnitude) bits. The complement of binary pattern 000 0111 becomes 111 1000. After we concatenate signed bit 1 and magnate bit string 111 1000, a 1 is added to the concatenate string 1111 1000. The final binary pattern after this converting process is $1111\ 1001_B$.

Real Number Representation in Digital Systems

In mathematics, we define real value as a noninteger value; that is, a value consisting of an integer part and potentially a fractional part. For example 3.625 is a real number in which 3 is its integer component and 625 is its fractional component. A decimal point is placed in between the two to separate them. To represent a real value in a digital system, we should reshape the floating point to be similar to *radix point* format but with some extensions. The standard format is called *normal form*. The normal form representation for a real number can be expressed as $f \bullet \text{base}^{E}$, where f is called *mantissa* and E is the *exponent*. Note that the mantissa needs to be normalized, so the digit after the separation point is nonzero. If needed the mantissa should be shifted appropriately to the right to the point where

Table 12.2 *Negative Value Representation Using Signed Magnitude Approach. The Drawback of Signed Magnate Representation is the Problem in Dual Zeros, +0 and −0*

Representation	Value Representation
0 0000	0
0 0001	1
0 0010	2
0 0011	3
0 0100	4
0 0101	5
0 0110	6
0 0111	7
0 1000	8
0 1001	9
0 1010	10
0 1011	11
0 1100	12
0 1101	13
0 1110	14
0 1111	15
1 0000	−0
1 0001	−1
1 0010	−2
1 0011	−3
1 0100	−4
1 0101	−5
1 0110	−6
1 0111	−7
1 1000	−8
1 1001	−9
1 1010	−10
1 1011	−11
1 1100	−12
1 1101	−13
1 1110	−14
1 1111	−15

the first digit (after the fractional point) is not zero. Then the exponent is adjusted properly to compensate for the shift. The normal form for 3.625 is $0.3625 \cdot 10^1$. In this decimal case, the base is 10 and the mantissa is 0.3615 and the exponent is 1. Let's explore an example and we'll have better

$$
\begin{array}{rl}
7_{10} \Rightarrow & 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1_{b} \\
\text{Compliment} \Rightarrow & 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0_{b} \\
+ & 1 \\
\hline
-7_{10} \Rightarrow & 1\ 1\ 1\ 1\ 1\ 0\ 0\ 1_{b}
\end{array}
$$

FIGURE 12.4 *A negative decimal integer −7 is equivalent to bit pattern 11111001.*

picture to understand how a computer system can use an 8-bit binary string to make up a floating number. Given a binary string 11010110_2, the representation protocol from left to right bit is defined as follows:

- **s1.** MSB (the first bit on the left) is used for the sign of the mantissa (1: negative, 0: positive)
- **s2.** The next four bits are for the magnitude of mantissa
- **s3.** The sixth bit is the sign of exponent (1: negative, 0: positive)
- **s4.** The remaining two bits on the right are for the magnitude of exponent

$$\underbrace{1}_{S1}\ \underbrace{1010}_{S2}\ \underbrace{1}_{S3}\ \underbrace{10}_{S4}$$

Based on this scheme, we can decode this bit pattern into its normal form.

- **s1.** MSB is 1, therefore the sign of mantissa is negative.
- **s2.** The normalized mantissa magnitude is 0.1010_2 in binary, which equals the value of 0.625 (=1/2 + 1/8).
- **s3.** Sign of exponent is negative.
- **s4.** The exponent magnitude is 10_2, which equals 2 in decimal value.

And, we actually know that the base must be 2 because it is a binary string. After we summarize all this information, the bit string can find its numerical value as:

$$-0.625 \times 2^{-2} = -0.625/4 = 0.15625$$

Converting real numbers to binary string numbers will often result in round-off errors. These can be important unless there is substantial bit depth. That is, the length declared for the bit pattern being used to accommodate floating point numbers needs to be long enough to make the round-off errors very small. Otherwise the magnitudes of mantissa and exponent will be truncated and drop precision. Consider making a calculation in which a formula is used that has the difference between two computed numbers in the denominator. If truncation causes these two numbers to become very close to each other, or disastrously equal to each other, the value of the function will approach infinity. In digital photography and image processing, where many complex formulae are used to process images, it is often valuable

Table 12.3 *Ranges of Numeric Value are Represented and Number of Bytes are Used for Various Data Types in Computer System*

Data Type Name	Bytes	Range of Values
signed int (signed __int32)	4	−2,147,483,648 to 2,147,483,647
unsigned int (unsigned __int32)	4	0 to 4,294,967,295
char (signed _int8)	1	−128 to 127
unsigned char (signed _int8)	1	0 to 255
signed short int (signed _int16)	2	−32,768 to 32,767
unsigned short int (unsigned _int16)	2	0 to 65,535
signed long int (signed __int64)	8	−9,223,372,036,854,775,808 to 9,223,372,036,854,775,807
unsigned long int (unsigned __int64)	8	0 to 18,446,744,073,709,551,615
float	4	3.4 $e^{\pm 38}$ (7 digits)
double	8	1.7 $e^{\pm 308}$ (15 digits)

to operate with a 16-bit per pixel image, even though a person cannot see more than can be represented in an 8-bit image. The difference is that round-off errors will be small compared to the required output values.

A question generally asked is how a computer knows that 1111 1001 is −7 and not +250 or even a floating number 1.875. In fact, a computer system does not know how to interpret 1111 1001 if we, as a user or programmer, do not specify its data type. In other words, it is the users' responsibility to predefine or declare an appropriate data type to the variable; therefore, the computer can process binary conversion with correct data representation. Table 12.3 shows the data type declaration and byte size and the range of values with respect to each data type.

ANALOG VERSUS DIGITAL SIGNALS

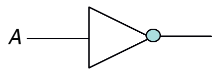
Analog signals have a physical value and no associated discrete numbers, whereas digital signals are comprised of numbers and not physical quantities. Analog signals can be converted to digital values and digital values can be converted to analog signals, which is how real world information in the form of analog signals are brought into a computer and how the results of computations are made to interact with the physical world. Once the information in a signal is rendered in a series of discrete levels, it can be quantized and encoded as a binary pattern. This concept and some issues about quantization will be discussed in the next section, where we will study how an analog signal is converted to discrete form, and a digital signal and its inverse can be converted into a real world output. In other words, how the brightness and color of a point on an object can be converted to a number, and how that number can be made to yield a particular brightness and color on a computer screen. We will examine the most basic digital circuit elements for doing this and explore how a digital system uses electric signals to represent and manipulate those binary values.

Boolean Algebra and Logic Diagrams

In 1847, English mathematician George Boole published the *Mathematical Analysis of Logic*, which eventually served as the basis for the development of digital computers. The form of algebra he invented considered only two values (1 or 0) for variables and certain basic functions which together are called *Boolean algebra*. Its graphical representation is called a *logic diagram*. This carries out algebraic functions and operations that can be appropriately used in designing the layout of logic circuit. The basic elements or building blocks used in logic diagrams are *gates* (sometimes called *logic gates*) and each of these performs one logic function. Each gate accepts a single binary input value denoted as A, or multiple binary input values denoted as A and B, and generates a signal binary output denoted as $X = f(A)$ or $f(A,B)$, where f denotes the function performed on the inputs. The set of outputs that result from the various inputs are set forth in a *true table* for the given logic gate. This lists all possible input combinations and the corresponding output. Simply said, the output value can be determined by gate type and the input values. There are six types of basic gates but they can be combined into rich arrays to form complex digital circuits that perform complicated arithmetic operations. The six gates are:

1. NOT gate (the NOT function is denoted by ~): Accepts one binary input value A and inverts it to form the output value X. The complement operation such as flipping a particular bit from 0 to 1 or 1 to 0 is implemented by using a NOT gate.
2. AND gate (the AND function is denoted by •): Accepts two inputs, *A* and *B*, and produces a single output, X, based on AND operation. The AND function is similar to the product of the two inputs. The output can be 1 only when both inputs are 1, otherwise the output is 0 when both or one of the inputs are 0.
3. OR gate (the OR function is denoted by +): The OR gate is very similar to the AND gate, which has two inputs and one output based on OR logic operation. The OR function is related to the addition operation. The output is always 1 when one of inputs happens to be 1. The output can be 0 only when both inputs are 0.
4. XOR gate (the XOR function is denoted by ⊕): Called *exclusive OR*. The property of XOR is interesting. When two input values are the

Logic Diagram



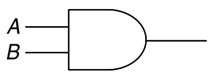
$X = f(A) = \sim A$

True Table

A	X
0	1
1	0

FIGURE 12.5 *Logic diagram, Boolean function expression, and truth table of NOT Gate.*

Logic Diagram



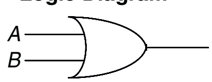
$X = f(A, B) = A \bullet B$

True Table

A	B	X
0	0	0
0	1	0
1	0	0
1	1	1

FIGURE 12.6 *Logic diagram, Boolean function expression, and truth table of AND Gate.*

Logic Diagram



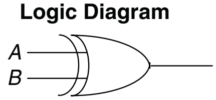
$X = f(A, B) = A + B$

True Table

A	B	X
0	0	0
0	1	1
1	0	1
1	1	0

FIGURE 12.7 *Logic diagram, Boolean function expression, and truth table of OR Gate.*

Logic Diagram



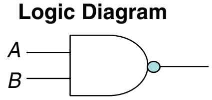
$X = f(A, B) = A \oplus B$

True Table

A	B	X
0	0	0
0	1	1
1	0	1
1	1	0

FIGURE 12.8 *Logic diagram, Boolean function expression, and truth table of XOR Gate.*

Logic Diagram



$X = f(A, B) = \sim (A \bullet B)$

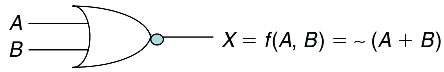
True Table

A	B	X
0	0	1
0	1	1
1	0	1
1	1	0

FIGURE 12.9 *Logic diagram, Boolean function expression, and truth table of NAND Gate.*

same (both are 0 or 1), the output is 0. That means XOR can be useful for making bit comparisons and signaling the difference. If two input conditions are not the same, XOR will yield a 1 as an output. Note that the XOR gate is different from the OR gate although both have

Logic Diagram



True Table

A	B	X
0	0	1
0	1	0
1	0	0
1	1	0

FIGURE 12.10 *Logic diagram, Boolean function expression, and truth table of NOR Gate.*

two inputs, similar diagrams, and produce a single output. When both inputs are 1, the OR gate's output is 1 whereas the XOR produces 0.

5 NAND gate (the expression of NAND gate is ~ ($\text{input}_A \bullet \text{input}_B \bullet \ldots$)): The NAND gate is derived from the combination of a NOT gate and an AND gate. The NAND gate is used essentially to produce the opposite output of an AND gate. Note that the NAND gate can also be derived by using NOT gates for both input values and their outputs are input to an OR gate. This derivation can be seen from the truth table.

6 NOR gate (the expression of NOR gate is ~ ($\text{input}_A + \text{input}_B + \ldots$)): The NOR gate is derived from the combination of a NOT gate and an OR gate. The NOR gate essentially is used to produce the opposite output of an OR gate. This can be seen from the truth table.

Combinational Circuits and Sequential Circuits

Circuits can be categorized into two major groups: *combinational circuits* and *sequential circuits*. For combinational circuits, the output is determined by the combination of input logics whereas in a sequential circuit, the output value is a function of input values and the current state of circuit. The expression of a circuit can be represented by the notation of logic diagrams, truth tables, and Boolean expressions. The solution of constructing gates with the identical output is not unique and it demonstrates *circuit equivalence*. For example, consider the combinational circuit in Figure 12.11.

The Boolean expression is $(A \bullet B) + C$, where the AND operation is applied to input values A and B. The output of operation $(A \bullet B)$ is used as input coupled with input C to an OR operation and produces the final output X. Let's use a truth table to express this combinational circuit, as shown in Table 12.4.

Let's try another circuit example, $(A + C) \bullet (B + C)$. This circuit consists of two OR expressions; the output of individual OR is the input of an AND operation. The logic diagram is shown in Figure 12.12 and the truth table is shown in Table 12.5.

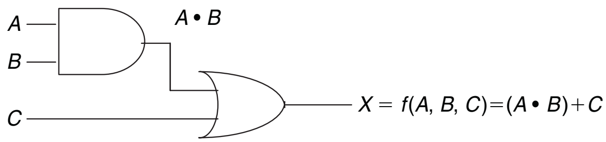


FIGURE 12.11 *Logic diagram and truth table of a combinational circuit (A • B) + C, which is a combination of an AND and an OR logic gate.*

Table 12.4 *The Truth Table of Combinational Circuit (A • B) + C*

A	B	C	(A • B)	(A • B) + C
0	0	0	0	0
0	0	1	0	1
0	1	0	0	0
0	1	1	0	1
1	0	0	0	0
1	0	1	0	1
1	1	0	1	1
1	1	1	1	1

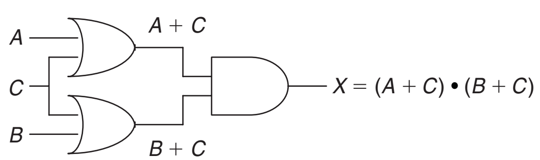


FIGURE 12.12 *Logic diagram of a combinational circuit (A + C) • (B + C), a combination of two individual OR gates whose outputs are input to an AND logic gate.*

Table 12.5 *The Truth Table of Combinational Circuit (A + C) • (B + C)*

A	B	C	(A + C)	(B + C)	(A + C) • (B + C)
0	0	0	0	0	0
0	0	1	1	1	1
0	1	0	0	1	0
0	1	1	1	1	1
1	0	0	1	0	0
1	0	1	1	1	1
1	1	0	1	1	1
1	1	1	1	1	1

Table 12.6 *Boolean Algebra for Circuit Equivalence*

Properties of Boolean Logic	Circuit Equivalence	
Identity	$A \bullet 1 = A$	$A + 0 = A$
Complement	$A \bullet (\sim A) = 0$	$A + (\sim A) = 1$
Commutative Law	$A \bullet B = B \bullet A$	$A + B = B + A$
Associative Law	$(A \bullet B) \bullet C = A \bullet (B \bullet C)$	$(A + B) + C = A + (B + C)$
Distributive Law	$(A + B) \bullet C = (A \bullet C) + (B \bullet C)$	

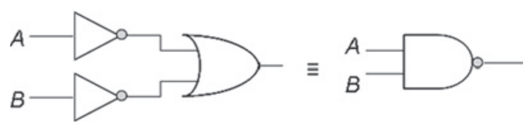


FIGURE 12.13 *DeMorgan Duality,* ~(A • B) = ~A + ~B.

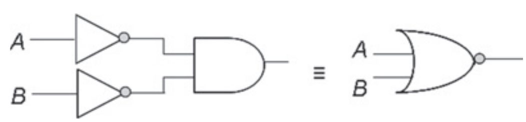


FIGURE 12.14 *DeMorgan Duality,* ~(A + B) = ~A • ~B.

We notice that the final result columns in both truth tables are identical. If two circuits produce the exact same output for each corresponding input logic combination, they are *circuit equivalents*. And the two examples in Figure 12.12 and Table 12.5 actually demonstrate *distributive law*, one of the important properties of Boolean algebra. Table 12.6 lists some properties of Boolean algebra for circuit equivalence.

There is an important Boolean algebra called *De Morgan's Theorem* proposed by Augustus De Morgan (1806–1871). De Morgan's Theorem consists of two relationships called *De Morgan Duality*:

$$\sim(A \bullet B) = \sim A + \sim B$$
$$\sim(A + B) = \sim A \bullet \sim B$$

This relationship states that the OR operation of two inverting input variables is equal to the NOT operator applied to the AND operation of the two variables. Inverting the output of an OR gate of two inputs is equal to inverting the value of individual variables first and taking AND operation. The logic diagrams of *De Morgan Duality* is shown in Figures 12.13 and 12.14.

De Morgan's Theorem is useful for circuit designs, and commonly used for simplifying circuit layout and design. Let's study the following example. Consider the following combinational circuit: $\sim(\sim(A + (B \bullet C)) + \sim(A \bullet \sim B))$. Now, we can apply Boolean algebra properties as well as DeMorgan's theorem to reduce this expression.

$$
\begin{aligned}
&\sim(\sim(A + (B \bullet C)) + \sim(A \bullet \sim B)) = \sim \\
&\quad (\sim(A + (B \bullet C))) \bullet \sim(\sim(A \bullet \sim B)) && \text{// DeMorgan's Theorem} \\
&\quad = (A + (B \bullet C)) \bullet (A \bullet \sim B) && \text{// Identity} \\
&\quad = A \bullet (A \bullet \sim B) + (B \bullet C) \bullet (A \bullet \sim B) && \text{// Distributive Law} \\
&\quad = (A \bullet \sim B) + 0 && \text{// Identity} \\
&\quad = A \bullet \sim B && \text{// Identity}
\end{aligned}
$$

The equivalent circuit $\sim(\sim(A + (B \bullet C)) + \sim(A \bullet \sim B))$ can be simplified as $A \bullet \sim B$, therefore the original five-gate circuit diagram is much reduced to a simplified two-gate version. Indeed, the entire $\sim(A + (B \bullet C))$ circuit and a NOR gate operation are redundant.

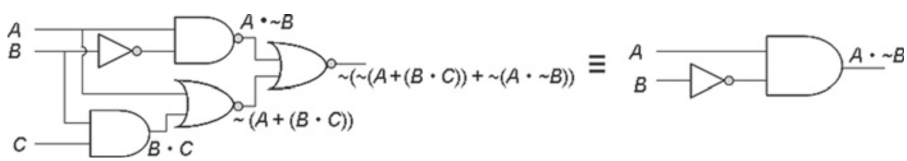


FIGURE 12.15 *Using DeMorgan theorem, A complex circuit* $\sim(\sim(A + (B \bullet C)) + \sim(A \bullet \sim B))$ *can be simplified to be* $A \bullet \sim B$.

Adders

A basic but important circuit design in digital systems is to perform the function of numeric addition. The arithmetic process of adding binary numbers has been introduced in the previous section; a type of combinational circuits called *adders*, which perform the function of addition, will be discussed in this section. Let's recall base-two binary addition. The result of adding two binary digits 1 and 0 is 1_B, and 1 plus 1 equals to 10_B, which produces a *carry bit*, 1. *Half adder* is a circuit designed to compute the sum of two bits as well as correctly produce the carry bit. To implement a half adder, we can first use a truth table to represent all possible outcomes of two individual outputs, Carry and Sum, from the summation of single-bit inputs *A* and *B*.

If we look at the output columns in Table 12.7, the results of output carry is the AND operation of inputs A and B, and the results of sum output actually is the XOR operation of A and B. Thus, the circuit diagram of half adder can be expressed as shown in Figure 12.16.

Table 12.7 *Truth Table of a Half Adder*

A	*B*	*A* + *B*	Carry	Sum
0	0	0	0	0
0	1	1	0	1
1	0	1	0	1
1	1	10	1	0

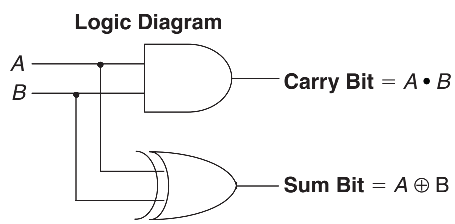


FIGURE 12.16 *The logic diagram of a half adder.*

Table 12.8 *Truth Table of a Full Adder*

A	*B*	Carry-in	Carry-out	Sum
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

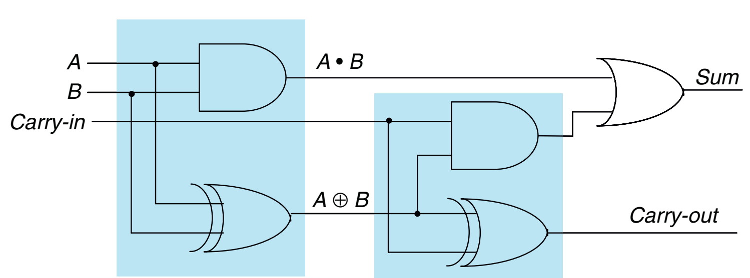


FIGURE 12.17 *The logic diagram of a full adder.*

A half adder can handle only the addition of two single digits but not enough to compute the addition of two multiple-bit binary values because a half adder does not have an additional input line "carrying in" a possible carry value. A *full adder* is the circuit that is extended from two half adders and can add the sum from the half adder to the carry-in bit. The output is a sum bit and a carry-out bit. A full adder circuit can simply be duplicated multiple times for multiple-bit values by simply connecting the carry-out from the lower-bit adder to the carry-in of the next higher-bit adder. Figure 12.17 illustrates the logic diagram of a full adder as well as the truth table.

Flip-Flop Circuit

Digital circuits can feedback the output of a circuit from its current state as the input of the same circuit for the next cycle; that is, the output of the current state is part of the input of the next state. And it is generally synchronized by a clock. This type of circuit is called a *sequential circuit*, which has the property of retaining bit value or storing digital information if a circuit is designed in large-scale. The *R-S latch*, which can store only a single bit (0 and 1), is the simplest sample of a sequential circuit, also called a *flip-flop* circuit. Given two circuit inputs Reset (*R*) and Set (*S*), from the name flip-flop, we can immediately get a sense that outputs *Q* and *Q′* of a two-input in a sequent ial circuit actually are always complements of each other.

Flip-flop circuits can be designed with either NOR gates or NAND gates. Using a NOR gate as an example, we know the output of a NOR gate is 1 only if both inputs are 0, but is 0 if any input is 1.

Assume that the input *S* (set) is 1 and input *R* (reset) is 0. The second gate's output *Q′* must be 0. Since *Q′* sends its output value 0 back to the first

Table 12.9 *A Truth Table Showing the State Transition of a Flip-Flop Using NOR Logic Gates*

R	*S*	*Q*	*Q′*
0	1	1	0
0	0	1	0
After Set is 1, Reset is 0			
1	0	0	1
0	0	0	1
After Set is 0, Reset is 1			
1	1	0	0

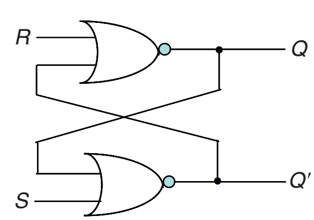


FIGURE 12.18 *The logic diagram of a flip-flop based on NOR gate design.*

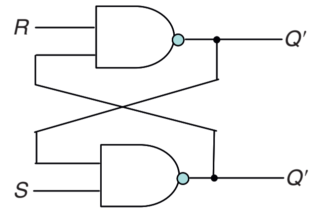


FIGURE 12.19 *The logic diagram of a flip-flop based on NAND gate design.*

Table 12.10 *A Truth Table Showing the State Transition of a Flip-Flop Using NAND Logic Gates*

R	*S*	*Q*	*Q′*
0	1	0	1
1	1	0	1
After Set is 1, Reset is 0			
1	0	1	0
1	1	1	0
After Set is 0, Reset is 1			
0	0	1	1

gate, and the input *R* (reset) value is also 0, the first output *Q* produces value 1. When output Q becomes 1 and the input *S* value returns to 0, the outputs *Q* and *Q′* remain the same. In the case of $S = 0$ and $R = 1$, value 1 in the reset input puts output *Q* to 0 and *Q′* to 1. When the reset input returns to 0, the same output remains. Table 12.9 shows the state transition.

Flip-flop circuits can also be implemented by using NAND gates. It follows the same manner as NOR gate design but outcomes are opposite.

When a 1 is applied to both the set and reset inputs in the NOR gate case or a 0 is applied to both the set and reset inputs in the NAND gate case, both Q and Q′ outputs turn to 0 in NOR gate design, or both outputs are 1 in NAND gate design. This situation indeed violates the property of a flip-flop that two outputs should be the complements of each other; therefore, these conditions must be avoided in setting or resetting input signals.

Multiplexer

A *multiplexer (MUX)* is a combinational circuit designed to select a single value from several binary inputs and direct it to an output line. Actually, the process of multiplexing is to transmit a large number over a smaller number of channels, such as a limited circuit line. Therefore, the multiplexer

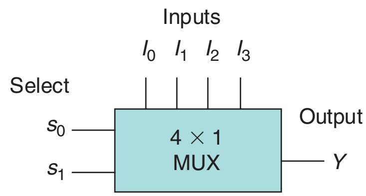


FIGURE 12.20 *A 4-to-1 multiplex (4 × 1 MUX) circuit block diagram. 4 × 1 MUX requires 2 selection lines s_0 and s_1 to select one of four inputs I_0, I_1, I_2, and I_3.*

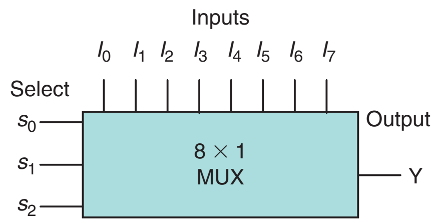


FIGURE 12.21 *An 8-to-1 multiplex (8 × 1 MUX) circuit block diagram. 8 × 1 MUX requires 3 selection lines s_0, s_1 and s_2 to select one of eight inputs I_0, I_1, …, I_7.*

Table 12.11

S_0	S_1	Y
0	0	I0
1	0	I1
0	1	I2
1	1	I3

consists of three basic components: a large number of inputs $\{I_0, I_1, \ldots, I_m\}$, an output line Y (carries the assembled value), and a smaller set of selection lines $\{s_1, s_2, \ldots, s_n\}$. In general, the number of inputs, $m = 2^n$, requires n selection lines. A 4-to-1 multiplex (4 × 1 MUX) means a multiplex has four inputs and one output and requires two selection lines. An 8 × 1 MUX then requires three selections. Figure 12.20 shows the circuit block diagram of a 4-to-1 multiplexer and Figure 12.21 shows the circuit block diagram of an 8-to-1 multiplex design.

A multiplexer indeed is a *decoder* circuit that decodes selection lines and sends the selected input to output Y. Table 12.11 shows the truth table of 4-to-1 multiplexer based on the circuit block diagram in Figure 12.20. Note that the value of MUX output Y equals to an input selected by the combination of two selection line s_0 and s_1. Figure 12.22 shows the detailed logic diagram for 4-to-1 multiplex design, which is often simplified by a rectangular block diagram as shown in Figure 12.20.

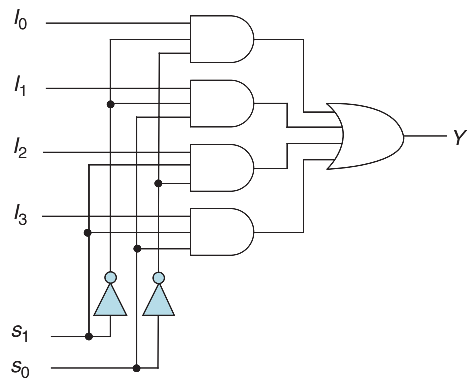


FIGURE 12.22 *A 4-to-1 multiplex logic diagram which is the detailed combinational circuit design equivalent to the 4 × 1 MUX simply circuit block representation in Figure 12.20.*

SUMMARY

The foregoing material shows that the well-known zeros and ones are chosen in conjunction with the digital circuitry described in the second portion of the chapter. More than two levels could have been selected, but the circuits work better with two, so digital technology has centered on binary data and circuitry that relies on two states. This combination results in a significant reduction in random noise in systems and is very stable. The numbers are based on two states and the circuits are based on two states, but they are both used in complex combinations and can produce excellent system performance. This is to say that zeros and ones result in reliable systems with the ability to solve an extremely wide range of logical and mathematical problems.

CHAPTER 13

File Formats and Compression

Digital images become useful when they are converted into a specific format that can be processed in a variety of applications. In order to make images "portable," they have to be saved and communicated in a standardized organization called a format. The more widely used formats are called standard formats. This enables either the same computer or a different device to successfully open the image, interpret the image content, encode or decode the image for viewing, and facilitate certain capabilities (e.g., enhancement) and inhibit others (e.g., unauthorized viewing).

RASTER FORMAT VERSUS VECTOR FORMAT

A graphics format is normally categorized into one of two major categories according to how the graphical data is stored and displayed. The first type of graphical representation is *raster graphic format*, usually called *image*, whose contents are described based on a regular array of raster pixel values. The second type of representation is the *geometry-based graphics*, also called *vector format*. The geometry-based format describes graphic contents based on a set of abstract geometric standard descriptions of lines, curves, and shapes, such as squares and circles, as opposed to the list (pixel array) of color information that raster-based graphics (images) use. We usually simply name the vector representation as *graphics* and call raster graphics as *images*.

The geometry-based graphics are generally more object-oriented; that is, the form of describing the geometric graphics is made up of a set of defined *line elements*. Every part of a digital layout of an image is made of lines and curves, and colors are applied both to the linear elements and the spaces they enclose. For example, the geometric shapes for a point in a Cartesian coordinate system—horizontal x axis and vertical y axis—can be represented by point locations (x, y). A solid white line (x_1, y_1) to

(x_2, y_2) can be presented by a vector using syntax with specific descriptions and parameters.

$$\text{line}(x_1, y_1, x_2, y_2,\ \text{thickness, 1, line\_type, 0, color, white})$$

Figure 13.1 shows the same line drawn from (1, 1) to (6, 6) but in two different graphics types. The left figure is a raster-based format and the right figure is based on vector format.

The two types of graphics format are often exchangeable. To translate the geometry-based graphic into a raster-type image requires the *rasterization* process, which calculates pixel color values by sampling the underlying geometric form. Rasterization is a process of vector-to-raster conversion and it does *point sampling*, which computes the brightness of the geometric shape at each pixel location. However, the rasterization process often needs to face the conventional sampling problem where image *resolution* and sampling accuracy are diminished due to round-off integer error for mapping image pixel coordinates. For example, the image aliasing effect that results in jagged stair-like patterns is observed on the edge of an image element after rasterization, a shown in Figure 13.2. The lower the sampling rate, the more prominent the steps. An image enhancement process called *anti-aliasing* can improve an image aliasing problem. Anti-aliasing uses neighboring pixels' values to reflect the proportion of the area covered by the line and to make the graphic more visually presentable.

The vector format is more favorable from an artistic point of view since it can deliver better visual presentation and can easily be scaled up and down in size without degrading. And, because of its abstract description in graphical content, it can achieve smaller file size for storage.

Raster format is easier to use for data from both an input device, such as digital camera, camcorder, scanner, and an output device, such as an inkjet

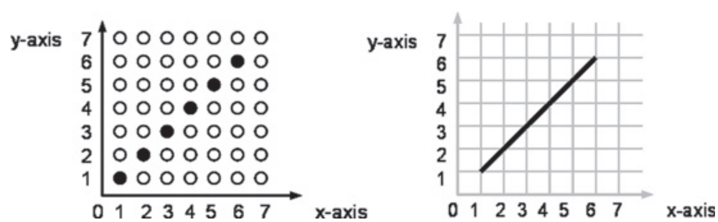


FIGURE 13.1 *Raster-based and vector graphic format for a line from coordinate (1, 1) to (6, 6).*

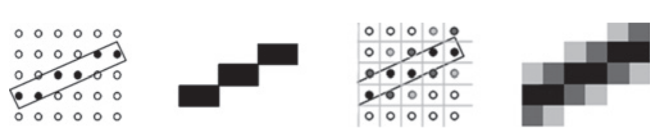


FIGURE 13.2 *Image aliasing effect that results in jagged stairs-like patterns on the edge (left). Anti-aliasing can improve image aliasing problem (right).*

printer and monitor. Note that laser printers using postscript language and/or other proprietary printer languages are really vector format devices. For output from or to devices, raster data is faster than vector format, because a vector-to-raster conversion does not have to be performed. In addition to fast input/output, raster format also provides an easier basis for image processing at the pixel level. This includes processes such as image filtering, transformation, enhancement, and other arithmetic processes.

COMMON COMPRESSION TECHNIQUES

Due to the advance of image acquisition devices that provide fast frame grabbing rates and produce higher resolutions, the amount of image data generated may be very large, and this results in difficulty in storage, processing, and communication. For example, the FBI and its largest division the Criminal Justice Information Service (CJIS) released a specification that regulates the quality and the format for acquiring both latent and live-scan fingerprints. There are some 200 million historical fingerprint cards on file and the FBI is still collecting some 30,000 to 50,000 new cards per day. The fingerprint resolution based on FBI compliance is at least 500 dpi, which results in about 10 MB per card. So the storage requirement for existing fingerprint images are 2000 terabytes. This number continues to grow quickly. To add to the problem, there are several large fingerprint data acquisition efforts that are ongoing at the Department of Homeland Security. Without image compression, it is impossible to archive those data.

Image compression tackles the problem of reducing the amount of data by minimizing the number of bits required to represent an image. The underlying basis of file size reduction is to remove redundant data while the original information can still be fully reconstructed or approximated at a later time. An image compression system must consist of two components: an *encoder* and a *decoder* for the *compression* and *decompression* functions, respectively. If data decompression can reconstruct the compressed data into an image that is an exact replicate of the original, it is called *lossless compression*. If data decompression cannot reconstruct the compressed data into an image that is an exact replicate of the original, it is called *lossy compression*.

Some general compression techniques that have been used commonly in information coding for communication purposes can be used for image compression as well.

Huffman Coding

Huffman coding is an entropy-based coding method and it provides lossless compression. The algorithm first arranges the order of probabilities of symbols and forms a symbol tree by combining the lowest probability symbols

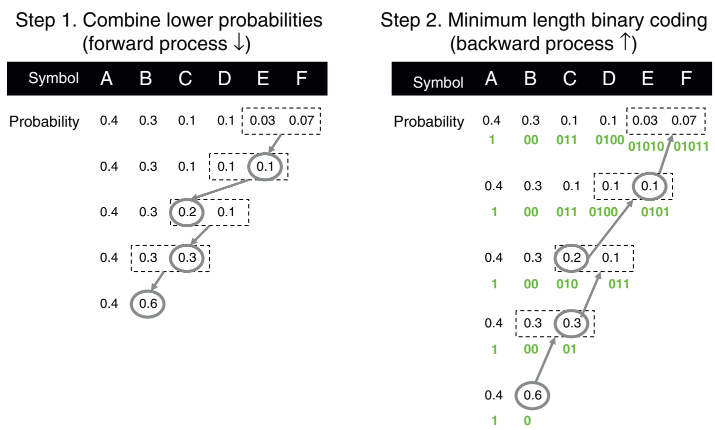


FIGURE 13.3 *The proces of Huffman encoding.*

into a single symbol. The second step is to start with the smallest source and trace back to the original source. Along the tracing back path, the minimal length binary code will be assigned to different probabilities paths. Figure 13.3 shows the procedure of how the Huffman encoding process works.

Assume six symbols/values {A, B, C, D, E, F} are used in a file, and the frequencies at which each symbol appeared in the file in terms of probability is 0.4, 0.3, 0.1, 0.1, 0.06, 0.04. We can construct a tree based on the step discussed earlier.

We can generate a *block code* with respect to symbols, where each derived binary block uniquely corresponds to each symbol:

A = 1
B = 00
C = 011
D = 0100
E = 01010
F = 01011

Then, given an encoded binary string, 010100111100, we can segment this string based on this block code and reveal the message to be ECAAB:

$$\underbrace{01010}_{E}\ \underbrace{011}_{C}\ \underbrace{1}_{A}\ \underbrace{1}_{A}\ \underbrace{00}_{B}$$

Run-Length Coding

Run-Length Coding (RLC) is a left-to-right encoding of a source string and relies on the number of times the same symbol appears in sequence.

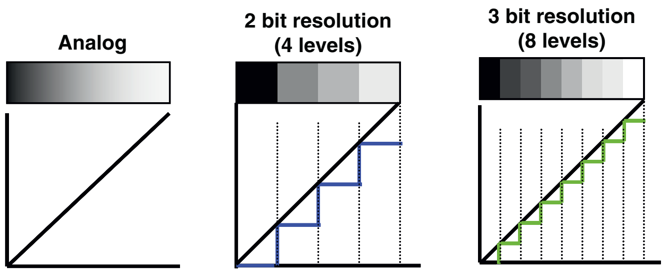


FIGURE 13.4 *2 bit (4-level) and 3 bit (8-level) quantizer.*

In binary, the symbols will be either 0s or 1s. RLC is a lossless compression method and the process is straightforward. This method has been used widely in facsimile (FAX) coding. For example:

A text string **aaabbbcccbbccccc** can be coded as 3a3b3c2b5c

and,

A binary string 11111111110000000111111111111111
can be encoded as 11one7zero14one.

Quantization (lossy)

Quantization is a process of approximating a set of continuous signals in a discrete vector. The original data inputs in a *quantizer* and the output always fall into one of a predefined finite number of levels, called bins. The quantizer is a functional module designed to make input values discrete and to represent the original signal with minimum loss or distortion.

Using uniform distribution to quantize input is a simple way to map input values into smaller and finite sets of bins. The quantizer using a uniform partition is called uniform quantizer, in which the input range is divided into levels of equal size. A uniform quantizer can easily be specified by its lower bound and the step size. Other types of statistical distribution can be used to design a nonuniform type of quantizer and the discrete output signals can possibly provide a better approximation to the original signal. Figure 13.4 shows the output of quantization for an analog input signal using a uniform quantizer with 2-bit (4-level) and 3-bit (8-level) bins. The finer bin partition provides better image fidelity but requires higher sampling frequency.

LZW

Lempel-Ziv & Welch (LZW) is a lossless compression and named after its developers, A. Lempel and J. Ziv, with later modifications by Terry A. Welch. LZW compression is used widely in GIF image files and as an option in

TIFF and PostScript file format. It is not always accepted in all image processing programs. LZW compression requires the association of a *code table* and generally this code table uses 12-bit codes that can provide 4096 entries. The table is built during the compression phase and the decompression can be achieved only by scanning each code from the compressed file and then finding and applying the corresponding translation from the code table. Normally code numbers from 0 to 255 are reserved for single bytes (SB) representation in the input file. However, it is optimal to combine many single bytes (*SB*) patterns into longer multiple byte patterns (*LP*). The longer patterns (*LP*) can be discovered by scanning the input file during compression. Figure 13.5 shows the LZW compression flowchart.

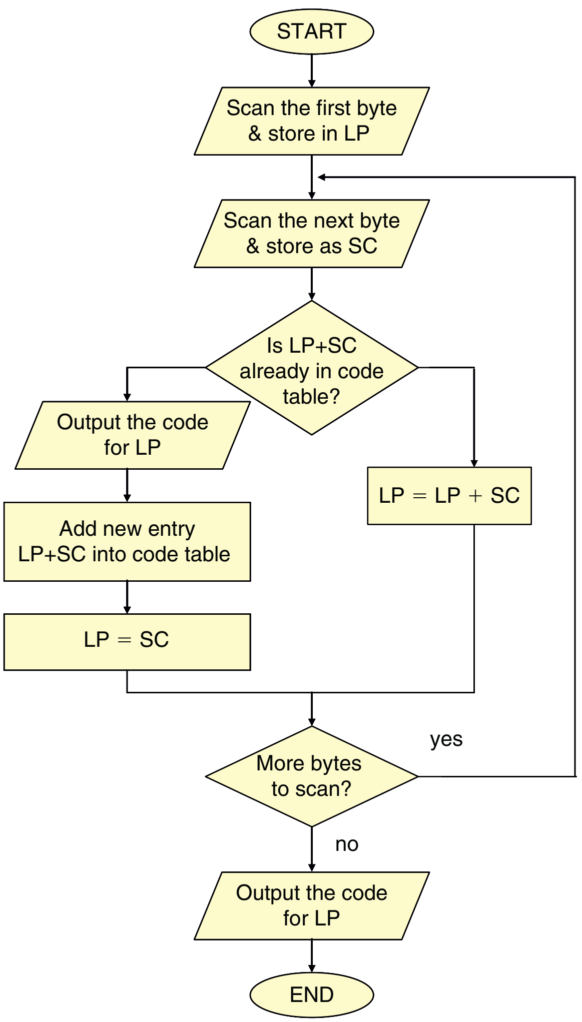


FIGURE 13.5 *LZW compression.*

When the LZW program starts to encode a file, the code table contains only single byte patterns in the first 256 entries (0~255). Entries after 255 of the table are blank. As the encoding continues, the LZW algorithm identifies repeated sequences in the data, and adds them to the code table. This process is repeated if there are more pixel bytes to scan, otherwise the process stops and the code will be assigned to LP. LZW can give about 2.5:1 compression rate but LZW is most effective when compressing indexed colors image format (which will be discussed in the next section), and is less effective for 24-bit RGB color images.

Although lossy compression causes a certain degree of information loss, a great advantage of lossy methods over lossless methods is that a lossy approach normally produces much higher compression rates. That is, the file sizes are considerably smaller than with lossless compression. The remaining question is, can the image still satisfy the requirements of the designated use? Some image file formats, such as RAW, GIF, and TIFF format in the cameras, use a lossless compression set by the manufacturers, whereas JPEG, WSQ, or some versions of TIFF use lossy methods.

High-resolution images require large amounts of storage space and would fill the storage media in your camera or a hard drive rapidly; therefore, compressing images has become a common practice, almost a necessity. There are basically three types of compression programs for images:

Lossless Compression:

- TIFF: Tagged Image File Format
- LZW (Lempel-Ziv-Welch) compression, recognized by all application programs
- The only lossless compression program on the market that is capable of working with continuous-tone images
- Compresses the image a maximum of about 2:1

Lossy Compression:

- JPEG and JPEG 2000-Joint Photographers Experts Group 2000
- Discards information as it compresses
- Compresses up to 30:1
- Compresses up to 80:1

Visually Lossless Compression:

- YCC-Kodak
- Kodak proprietary with Photo CD
- Compresses up to 5:1
- Removes visually unimportant pixels

COMMON STANDARD IMAGE/GRAPHIC FILE FORMATS

A standard file format is necessary and important since it allows the movement of images between applications. Sometimes file compression is added on to save disk storage space and increase file transmission speed during data communication. As discussed in the previous section, the vector-based format uses geometrical description and a piece-wise line element representation to describe an image. However, in the raster format, where the image is represented by color pixel array, there are various types of array configurations that can be used to hold pixel color values. Two major image color types are RGB color format and indexed color format. Of course, there are many other color formats, such as Lab color for displays, or CMYK color format, which is mainly for color printing. But, the RGB and indexed color formats are the most commonly used. For example, the GIF image file is based on indexed color. JPEG and TIFF encode red, green, and blue (RGB) color channels.

RGB (red, green, blue) is a three-channel color space because it describes all colors as the combination of some percentage of its three primary colors: red, green, and blue. Indexed colors usually derive from RGB color. Instead of R, G, B pixel arrays, indexed colors maintain index color palette files called colormaps and an array that stores pointers linked to a colormap array. Index color palettes are usually smaller and more efficient for network transmission but require rendering for display. This is the reason why most Internet images are indexed color. Figures 13.6 and 13.7 illustrate both RGB colors and indexed color images.

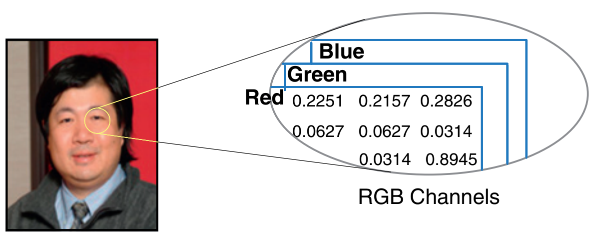


FIGURE 13.6 *RGB color image.*

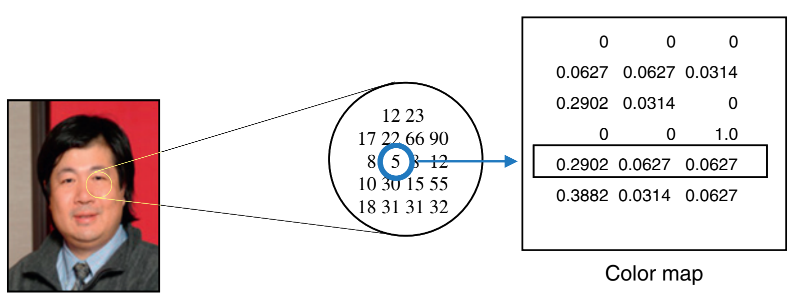


FIGURE 13.7 *Indexed color images.*

All image files have a unique format, and can be recognized by the extension attached to the file. Most digital cameras have the option of at least two formats and sometimes a third. The options usually include .JPG, .TIF, and RAW. Many manufacturers are removing the .TIF option and including only .JPG and RAW.

Common formats are:

- (TIFF) *Tagged Image File Format*. The most common format, because it is supported by all computer platforms and applications and is used by many digital cameras as an option. It also supports LZW compression, which is a lossless compression method.
- (JPEG) *Joint Photographic Experts Group*. Commonly used with continuous tone images to reduce file size and is an option on all digital cameras. JPEG discards data as it compresses the image and is considered *lossy*. Higher compression levels result in lower image quality. This is probably the most popular format in use with digital cameras.
- (JPEG 2000) *Joint Photographic Experts Group 2000*. JPEG 2000 is a new image coding system that uses state-of-the-art compression techniques based on wavelet technology. Its architecture should lend itself to a wide range of uses from portable digital cameras to advanced prepress, medical imaging, and other key sectors. Presently it is not supported by any digital cameras or applications. A plug-in can be purchased to add the capability to PhotoShop.
- (RAW) *Proprietary Camera Format*. A proprietary format established by camera manufacturers that records a raw image to the camera media card. Software capable of reading the format must be installed on the computer. PhotoShop is capable of reading most RAW formats.
- (GIF) *Graphics Interchange Format*. Only capable of reproducing 256 colors and is commonly used to display indexed color graphics and images on the World Wide Web. It is a compressed format that minimizes transfer time over phone lines. This is not an option on digital cameras.

There are many other file formats, but most are unique to the application. One you will discover as you use PhotoShop is the PSD format. This is unique to Adobe and cannot be imported into most other applications.

Tagged Image File Format

Tagged Image File Format (TIFF) is the industry standard for the exchange of image files. TIFF is designed with three major goals:

1. Extendability. Provide the ability to be compatible with older image formats in new image applications.

2 Portability. Perform equally well among various operating system environments, for example, Apple, Windows, and UNIX.

3 Revisability. Can be useable as an internal data form for image editing applications.

TIFF files consist of three data structures, as shown in Figure 13.8. The first data structure is called *Image File Header* (IFH), which is located at offset zero (the very beginning) of the file and its length is only 8 bytes long. This contains the information of byte order, version, and the offset to the first image file directory (IFD). The second structure is the *Image File Directory* (IFD) and it can be one or more directory entries. Each *Directory Entry* (DE), the third structure, is 12 bytes in length and is segmented into four fields, including a tag field that points to the actual raster data location.

Since TIFF retains the raster RGB data and maintains tag fields for image descriptions, the image files are usually large. Normally, it does not

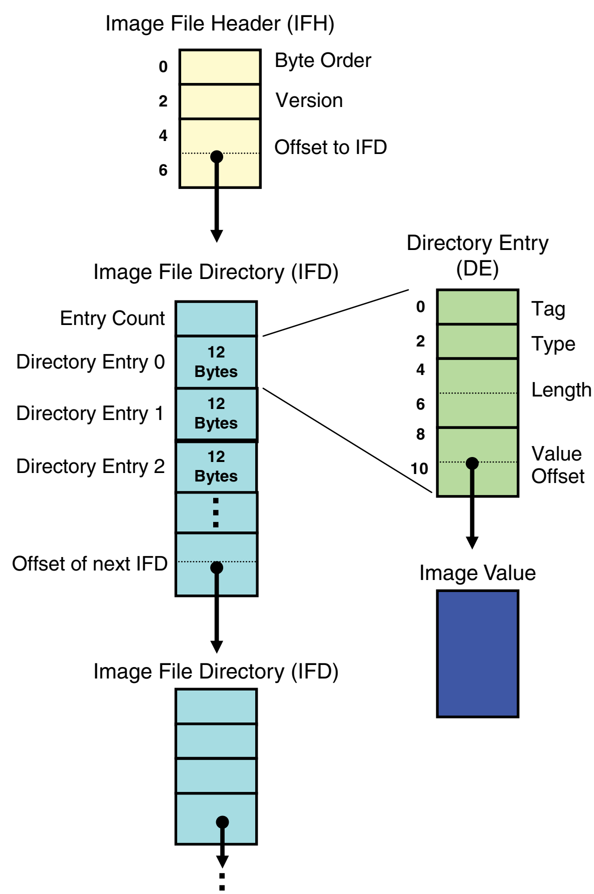


FIGURE 13.8 *Data Structure of TIFF file format.*

involve compression but LZW compression can be applied to generate compressed TIFF.

JPEG Format

The Joint Photographic Experts Group (JPEG), a joint committee between ISO and ITU-T (formerly CCITT), standardized image compression mechanisms to compress either full-color or gray-scale images of natural, real-world scenes. JPEG is a lossy compression; that is, the decompressed image isn't quite the same as the one you started with and the degree of loss can be varied by adjusting compression parameters—the quantization levels. Note that when the amount of quantization change applied is very small or nonexistent, the amount of loss is negligible and the amount of compression is relatively low. The file will still be considerably smaller than a TIFF, uncompressed version of the image. As more quantization is applied, the amount of loss increases and the file size reduction is increased as well.

Figure 13.9 illustrates the information loss in nominal JPEG compression. Although we cannot visually tell the difference between images (a) and (b), we can notice there are artifacts that occur nearby the edges of objects and there is some loss of high frequency information. Image (c) shows the difference between images (a) and (b).

The JPEG file encoding process, shown in Figure 13.10, consists of several steps:

1 **Preparation of an 8 × 8 micro block**. This description is based on the images commonly in video, but virtually the same process can be applied to any digital image. An NTSC image format consists of 640 × 480 pixels with 24 bits/pixel (3 colors with 8 bits per color). The first step of JPEG compression has to do with color space transformation. The representation of the colors in the image is

FIGURE 13.9 *(a) Original image, (b) JPEG image, and (c) the difference between image (a) and (b).*

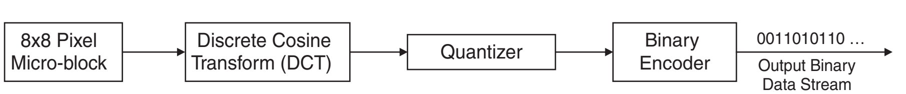


FIGURE 13.10 *JPEG file encoding process consisting of four major steps: 8 × 8 Micro Block, Discrete Cosine Transformation (DCT), Quantization, and Zig-zag Binary Encoding.*

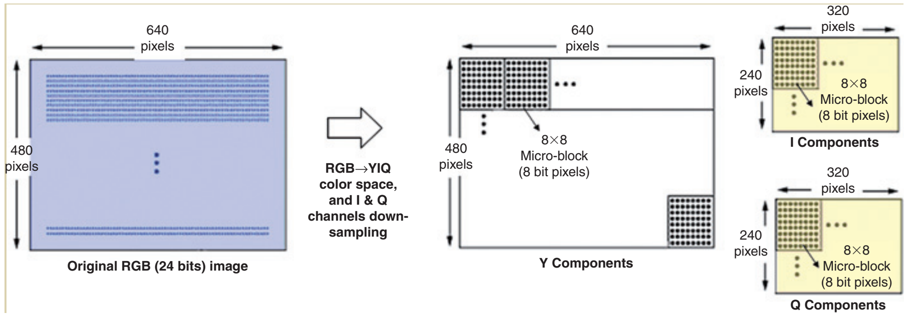


FIGURE 13.11 *8 × 8 pixel blocks segmented from Y, I, Q matrices.*

converted from RGB space to YIQ color space. YIQ is a type of color scheme that includes one luminance component (Y), which carries the brightness information, and two chrominance components (I and Q) for color representation. Y, I, and Q channels are placed in separate matrices, where Y is the size of 640 × 480. The resolution in the chrominance matrices is reduced by a factor of 2 so that the I and Q matrices have only 320 × 240 pixels each. This reflects the fact that the eye is less sensitive to color details than to brightness. Each matrix is then divided up into blocks of pixels that are 8 × 8. Therefore, there are 4800 (0~4799) blocks in Y matrix, and the I and Q matrices contribute 1200 blocks each. Figure 13.11 shows how 8×8 pixel blocks are segmented from YIQ space.

2 **Discrete Cosine Transformation (DCT)**. After micro-block preparation, the DCT is applied on the total 7,200 8 × 8 micro blocks, each treated separately. The output after DCT still remains an 8 × 8 matrix of DCT coefficients. DCT coefficients are the amplitudes of the cosines for each of the frequencies and indicate how much spectral power is present at each spatial frequency within the original micro block. Note that, at this point, DCT is lossless with the only losses being due to round-off of floating-points numbers, which will happen any time pixel data are processed mathematically. Figure 13.12 shows the DCT coefficients in an 8×8 block of the Y matrix.

3 **Quantization**. The amplitudes of the frequency components are quantized. Since the human eye is most sensitive to low frequencies (upper left corner) components and less sensitive to high frequencies (lower right corner), the magnitude change in high-frequency components can be quantized by using fewer bins. The trade-off is to sacrifice accuracy in less important coefficients to achieve a higher compression ratio. The numbers in the quantization tables can be scaled up (or down) to adjust *quality factor*. Keep in mind that the quantization is the lossy part in JEPG compression whereas the

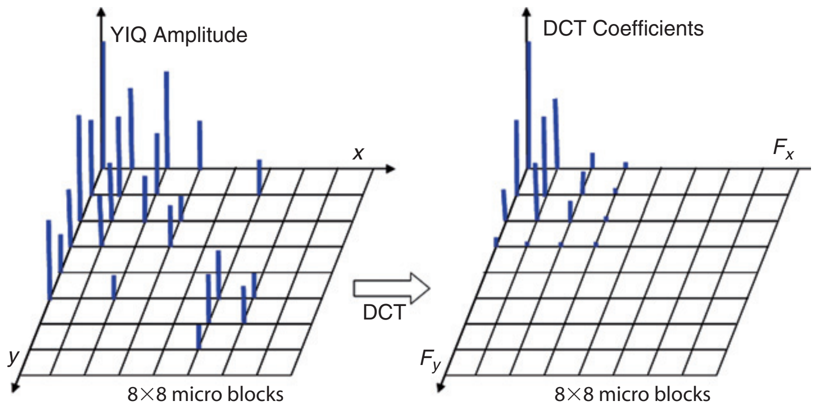


FIGURE 13.12 *DCT coefficients.*

previously introduced DCT and next step binary coding are lossless components.

4 **Zig-zag binary encoding**. The resulting data, after quantization, for all 8 × 8 blocks is further compressed using a lossless algorithm, zig-zag run-length binary encoding. This takes the square array of numbers and converts it to a single string of numbers from the matrix. Each number is next to the number that was its diagonal neighbor in the square. The algorithm first calculates the value of the difference between each current number and the previous one. Because of the geometry involved, the zig-zag arrayed numbers are more likely to be close in value to neighbors. Run-length encoding method for binary coding is used to encode nonzero components out of a 1 × 64 vector. JPEG compression can easily reach 20:1 compression ratio.

WSQ and JPEG 2000

Wavelet Scalar Quantization (WSQ) and JPEG 2000 both compress images based on wavelet transformation. Why is a new and better compression technique beyond JPEG needed? Referring to the FBI fingerprint example we saw earlier, if 10 segmented fingerprint images per card are digitized at the resolution of 500 dpi, that would be 10 images with 768 × 768 pixels at 256 gray levels, equal to 10 MB/card (=589,824 bytes × 10 fingers × 2 for both a dab and rolled fingers). For 200 million cards and without applying any image compression, it requires 2000 terabytes of storage space. Normally, a single image takes three hours to be transmitted via the police commonly used 9600 baud channel. You can imagine how long it will take for a police officer to conduct fingerprint matching. The FBI requested and innovated an image compression technique that targets the compression ratio of at least 0.75 bits per pixel (10.7:1) and they stated that JPEG image quality is unsatisfactory for fingerprint matching. In 1993, the Wavelet Scalar Quantization (WSQ) compression was developed by Tom Hopper at the FBI with joint effort from Jonathan Bradley and Christopher Brislawn at Los Alamos National Laboratory in supporting the FBI AFIS fingerprint system. JPEG 2000 was developed later on, based on WSQ compression, and it offers the most potential image compression algorithm in the industry today.

The following is a brief listing of sequential steps (codec) for WSQ:

1. The image is decomposed into a number of spatial frequency subbands (typically 64) using a Discrete Wavelet Transform (DWT). This provides multi-Resolution Analysis (MRA) capability. Figure 13.13 shows 2D wavelet transform for three-level subband coding to generate a multi-resolution quadratic tree using the Lena test image.
2. The resulting DWT coefficients are quantized into discrete values.
3. The quantized subbands are concatenated into several blocks, similar to the zig-zag process used in JPEG and compressed using a Huffman run-length encoding.
4. Decoding is to reverse the order. Start with Huffman decoding. Then decode the quantization. Finally, invert Discrete Wavelet Transforms (IDWT).

The images in Figure 13.14 show the difference of fingerprint image quality after JPEG (a) vs. WSQ (b) compression. Notice that the image quality in (b) is superior to image (a). Image (a) shows a block artifact in the center.

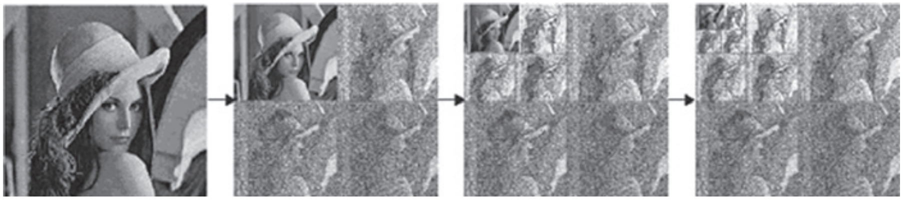


FIGURE 13.13 *2D Wavelet decomposition for Multi-Resolution Analysis (MRA).*

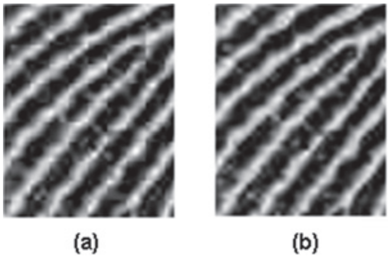


FIGURE 13.14 *Image quality comparison between JPEG (a) and WSQ (b) compression.*

JPEG 2000 can give better image quality than JPEG at any given compression level, and though JPEG is not normally used at ratios greater than 20:1, JPEG can achieve ratios of 80:1. JPEG shows remnants of the 8×8 micro blocks as artifacts and JPEG 2000 has a tendency to show a bit of blurring of very high frequency elements of an image. For reasons not clear to us, JPEG dominates the industry, with most digital camera manufacturers offering it on their cameras; JPEG is rarely seen and it is not available on cameras.

PSD Format

This is a proprietary format used by Adobe, Inc. in its software products. This format retains the layers and processing steps that might have been used while editing an image and is an excellent means for temporary storage of images that will be reedited in the near future. But, like any proprietary format, it is not recommenced for long term storage or for use by others who may not have a compatible editing or presentation software package.

RAW

RAW files do not hold viewable images. They contain the raw data from a specific camera and they need to be converted to one of the standard file formats before they can be viewed, enhanced, or edited. Each camera manufacturer provides proprietary software for converting RAW files from a given camera. The camera captures several items of information that are normally applied inside the camera in order to convert an image, and those information elements are included in the RAW file. The operator has the option of using those elements or different values as his or her choice. These choices can have profound influence on the resulting image. The process of computing the image affords the operator a bit of extra range of adjustment but takes quite a bit of time.

RAW files are smaller than TIFF files, but larger than unquantized (zero-compressed) JPEG files. They are not suitable for conversion by someone other than the photographer unless the camera-offered information elements are used since that individual did not see the original scene. RAW files are not suitable for archiving since the conversion software is proprietary. Typically it is good practice to convert RAW files into TIFF images and to save those.

CHAPTER 14

Sensor Chips

There are three types of sensor chips currently used in digital image capture devices. They vary in the arrangement and technology used to capture light from different portions of the image and how they then process that information. The two most widely used types are Bayer array charge coupled devices (CCDs), and Bayer array complimentary metal on silicon (CMOS) chips. One manufacturer uses octagonally arrayed pixels. For scanners, the pixels are aligned in a single row and there are four separate rows—one for each of the primary additive colors (and sometimes one for infrared). Video cameras might use a different array pattern but are generally similar. One manufacturer uses a different technology to stack the three color reception sites vertically. The Eastman Kodak Company recently has announced a new version of the Bayer array (Dr. Bayer was a Kodak scientist) that will give higher sensitivity and better dynamic range.

CAPTURING THE LIGHT

All the digital light sensors commonly used in digital imaging devices are the same in that they comprise small tiles, or picture elements, called *pixels*. All the pixels are part of a larger integrated circuit chip, referred to as a *sensor* chip or *imager* chip. It is usually made on a silicon-based crystal. A pure silicon crystal is composed of silicon atoms in a regular array. Each silicon nucleus has a fixed location in a regular matrix and the electrons of each atom serve to bind the nuclei together. Silicon has a valence of four, and so each atom is bonded to four surrounding atoms. In the absence of radiation, the crystal does not support the movement of electrical carriers, such as electrons. It is an insulator, much like most plastics, and the opposite of a conductor, such as copper. See the schematic in Figure 14.1.

In the case of silicon, when a photon of the right wavelength (remember that wavelength is inversely related to energy) is absorbed, it will energize an electron to the point of being set free from the matrix. Once loose, the electron

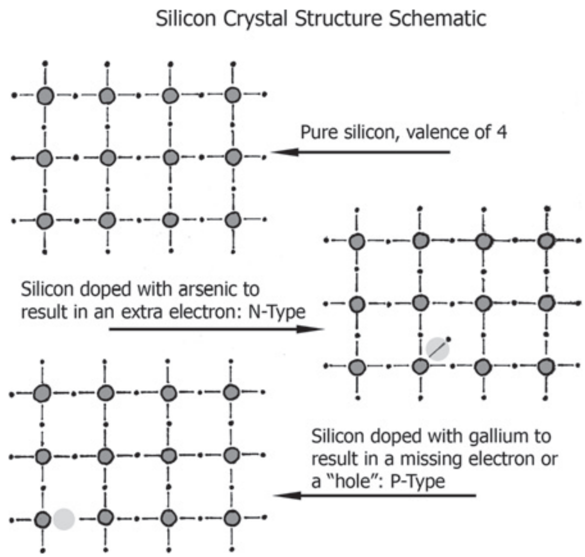


FIGURE 14.1 *Silicon Crystal Structure Schematic. Silicon has a valence of 4 and it shares electrons with the adjacent four silicon atoms. This structure does not make a good conductor of electricity since there are no mobile carriers of charge. If a few atoms of arsenic or other valence 5 atoms are inserted into the crystal, then there are unbound electrons. This material will conduct electricity using negative carriers, and is called an N-Type semiconductor. If atoms of gallium were inserted instead, there would be holes where there should have been electrons. These are able to move by a process resembling musical chairs and appear to be positive carriers. This is called a P-Type semiconductor. Placing P-Type and N-Type semiconductors in contact creates a p-n junction, which gives rise to transistors, diodes, and photo diodes.*

is relatively free to move around the crystal. The electron carries a negative electrical charge and is called a *negative carrier*. The movement of electricity in a copper wire is composed of the movement of electrons. In a silicon crystal, called a *semiconductor*, the release of an electron from its place in the matrix leaves behind a hole in the matrix. This is referred to as a *hole* and it, too, is somewhat free to move by a sequential substitution process—musical chairs. In the aggregate, the holes appear to be positive electrical carriers. If an electrical voltage is applied across the crystal, the negative carriers will move in one direction and the positive ones in the other. In summary, a silicon crystal is an insulator in the dark and a conductor in the light. The amount of charge that moves is equal to the number of photons absorbed and, therefore, proportional to the amount of incident light (not all the incident photons will be absorbed). The amount of charge released divided by the number of incident photons is referred to as the *quantum efficiency* of the crystal.

In P-Type semiconductors, an impurity is added which has a valence of three. The result is that the crystal has intrinsic holes. In an N-Type semiconductor, the impurity has a valence of five and so the crystal has intrinsically extra electrons. P-Type materials do not conduct electrons and N-Type materials do not conduct holes. If a P-Type crystal is joined to an N-Type one, current will not flow in the direction in which holes have to traverse the N-Type section and the electrons have to traverse the P-Type section. This is called reverse biased. If the device is forward biased, the holes and electrons will meet at the junction and annihilate each other and current will flow. The insertion of free electrons (and left behind holes) due to absorption of incoming photons will make the device conduct a photocurrent even when reverse biased.

Consider a sandwich made with thin metal plates instead of bread, and plastic for the filling. The metal plates are electrical conductors and are called the *plates*, and the plastic is an electrical insulator called the *dielectric*. If the plates were connected to a battery—one to each terminal—electrical charge would flow from the battery to the plates; and once the voltage across the insulator became equal to that of the battery, all action would stop. If the wires from the battery were carefully removed, the device would continue to hold the voltage, assuming the insulator did not leak at all.

This little thought experiment had electrical charge carriers flow to the metal plates and then stay there at a certain voltage. This is the description of a device called a *capacitor*. It holds electrical charge. The capacity of a capacitor is defined as the amount of charge it can hold per volt applied. The capacitance in farads is equal to the electrical charge in coulombs divided by the electrical pressure in volts. The capacitance is a measure of the physical device, whereas the volts and coulombs are in response to the outside world. But a strict, fixed relationship is maintained between the volts and coulombs. If you double the coulombs, the voltage will double. Each pixel is essentially a miniature capacitor, except that the dielectric creates charge when light is impinged. And the amount of charge is proportional to the number of photons per unit area. The result is that the amount of charge left on the plates of the pixel when the exposure is stopped is linearly proportional to the illumination. The result is both linear and analog—even in *digital* cameras. The basic light gathering process starts out as an analog signal. Once the exposure is complete (the camera's shutter is closed again), we can measure the voltage and infer the charge (coulombs) or measure the charge directly. In either case the result is to derive an electrical signal that is proportional to the incident light. At some point, this analog signal will be converted to a digital number.

In practical devices, the sensor elements are actually *photodiodes*. These create charge in a fashion somewhat similar to what was described, but prevent it from recombining. In the literature, pixels are based on photodiodes, or p-n junction diodes. But the basis is similar: light liberates charge, and the charge accumulates and creates a voltage. The analog voltage value is read and converted to a digital number.

In Chapter 2, performance characteristics of imagers at the extreme low and high levels of light were described, and it is useful to repeat a summary of that. At very low light levels, some charge will flow just due to the fact that the chip is at some finite temperature. So there is a low limit of charge, which is variable and not due to light. This is *noise*. At very high light levels, more charge is created than the capacitor can hold, and the excess spills over to other circuit elements. This is *saturation*. At light levels below the

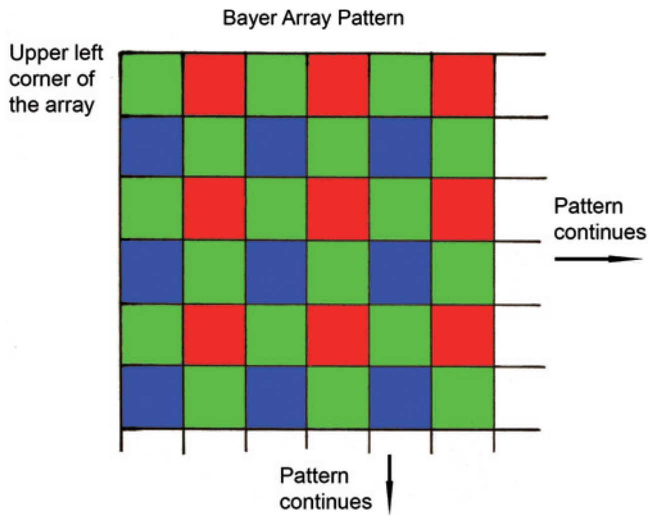


FIGURE 14.2 *Bayer Array Pattern. Most digital still cameras utilize the Bayer array pattern, which places either a red, green, or blue filter on top of each pixel. This pattern is designed to give extra weight to the luminance value, as carried by the green filtered pixels. The red and blue pixels basically carry chrominance information.*

noise limit and above the saturation limit, the photo sensor does not operate as a true indicator of light.

BAYER ARRAY

The Bayer array comprises a repeating rectangular array of pixels, with each one covered by a filter as shown in Figure 14.2. Each pixel is a light-sensitive portion of the main chip and is covered by a filter that passes either red, green, or blue light. Below the filters, each pixel is like all the others. The array pattern is what is referred to as the *Bayer array*. This pattern has every other pixel covered with a green filter. The green record is aligned most closely with the luminance of an image and is therefore the most significant contributor to the visual impression of sharpness. The green record must have the best resolution if the images are to appear sharp. The human eye also has an overabundance of green sensitive cones. Since every other pixel is green, half of all the pixels are green. The remaining half is divided equally into red and blue pixels.

The red pixels will be in one row interspersed by green, and the next row will have blue pixels interspersed with green. The same arrangement is maintained vertically. The result is that every blue pixel is surrounded by four green pixels and four red ones. Likewise, every red pixel is surrounded by four green pixels and four blue ones. All green pixels are diagonally surrounded by four other green pixels and at least two of each of the red and blue pixels.

The important properties of the Bayer array are that: (1) each pixel captures light from only one of the three additive primary colors of red, green, and blue; (2) the highest sampling of the image is done in the green portion of the spectrum; and (3) there is close proximity of each pixel color to the other two colors.

The digital signal coming from a Bayer array chip is a string of numbers in which the value of each number is an indication of the amount of light that the particular pixel captured, and its location in the string is an indication of where that pixel was in the original array. The color of the light that that pixel saw is inferred from its location in the original array. This raw information is just that—raw; it is not a viewable image. The viewable image must be computed from this raw data and certain other indicators from the camera. Looking at the string of numbers is no more like seeing an image than looking at the lands and troughs in the various rings of a music CD is like listening to Beethoven.

All pixels have certain inherent properties that are important considerations in the design of a sensor chip. First of all, the larger the area of the pixel, the more sensitive it is. It is responding to photons per unit area, and the bigger the area, the more photons it will capture, all other things being equal. All pixels have a certain amount of dark current. That is, they leak in the dark. The silicon is not supposed to generate carriers in the dark, but there is always some generation. This results in noise in the image in the form of a speckle pattern in the dark areas. The result is that chips with larger pixels have more signal and (usually) proportionately lower noise.

Inevitably there will be a few defects in a chip. This can result in pixels that are too sensitive or not sensitive enough, compared to others on the chip. Testing and conditioning of the signal can minimize the impact of these defective pixels. Also, the process of passing the charge to be read is not lossless and totally consistent. It is something like pouring maple syrup from a measuring cup into the mixing bowl. It leaves room for a bit of variability—that is, noise. Again this noise will be most evident at low light levels.

Silicon has a relatively low native sensitivity to the shorter wavelengths (blue) and a higher sensitivity to longer wavelengths (red). The result is that the dark current noise is more pronounced in the blue record. Fortunately people are not as sensitive to the blue record. But in order to keep a reasonable balance among the three records, the sensitivities of the green and red pixels are held back to be consistent with the blue ones. The end result is that the performance of the blue pixels tends to limit the overall performance of a chip.

Finally, the cost of a chip goes up with its area. Make the chip twice as wide and it will cost four times as much. This is due not only to the cost of the materials but is impacted by the incidence of defects on the wafer from which the chips are taken. If the wafer has, let's say, three defects, and a single chip is taken from that wafer, the chances are good that that chip will have at least two of those defects. If it were possible to work with chips that were half that size on each side, then we could get four chips from the same wafer. It is now almost certain that one, and possibly two, of the chips will be defect free. The result is that smaller chips tend to have a higher yield-per-wafer and are therefore cheaper to produce. This is in addition to the reduced cost of the basic material.

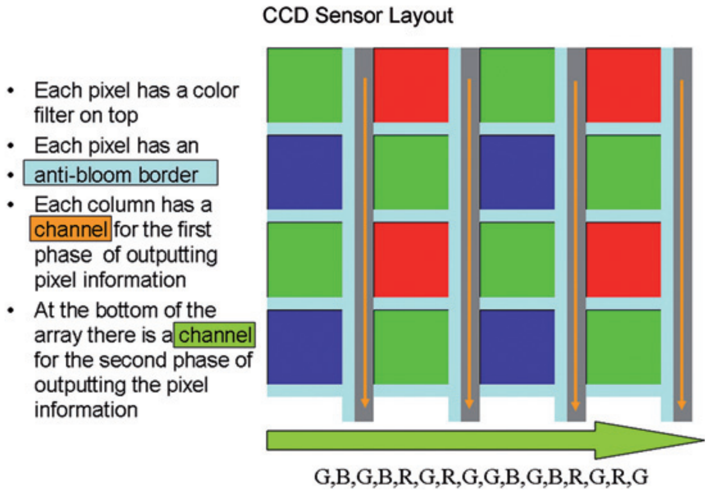


FIGURE 14.3 *CCD Sensor Layout. Most CCD sensor chips follow the Bayer array. There is an anti-blooming area surrounding each sensitive area. The charge generated within each pixel is moved to a hidden cell just to the side of each pixel. Then the columns of held charge are moved downward to a similar series of cells along the bottom of the chip. The charge packets move systematically out of the chip where they are converted to a digital number and the string of numbers is the output from the chip.*

Given that larger pixels are desirable from a sensitivity and noise reduction standpoint, but that smaller chips are cheaper to produce, there is a constant struggle between these two factors.

CCD AND CMOS CHIPS

CCD and CMOS chips have different structures but, performance-wise, they differ primarily in how the analog values from the pixels are converted to digital signals and how those signals are extracted from the chip. Except for battery life, most of the differences will not be noticeable to the user of a camera.

In the case of CCDs, there is a column of capacitors adjacent to each column of pixels (see the schematic in Figure 14.3). Once the shutter is closed, the charge from each pixel is moved from the pixel itself to the adjacent capacitor. This happens all at once. Then the charge from each capacitor in the column is transferred to the one just below it. The process is likened to a string of men shifting buckets of water along in a bucket brigade. The charge in the bottom-most capacitor is transferred to a similar capacitor that is one of many in a row that lines the bottom of the array. All the columns shift their charge down the chain and all the bottom-most ones shift into the row at the same time. In synchronization with the movement down the columns, the charge packets are moved along the row and taken out of the CCD proper. At this point there is an analog-to-digital converter (ADC) that takes the stream of analog signals in and sends out a stream of digital numbers.

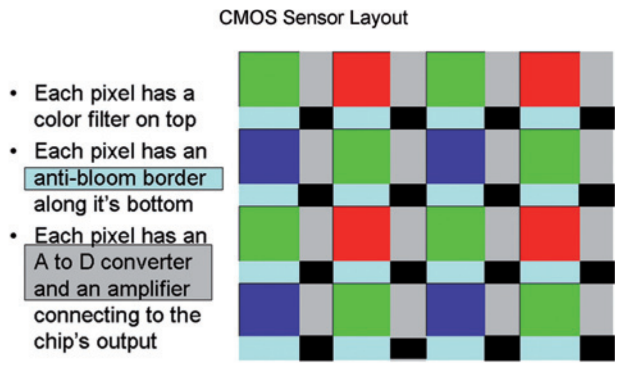


FIGURE 14.4 *CMOS Sensor Layout. These chips have light sensitive areas with a partial anti-bloom barrier. Also, each pixel has its own A to D converter and amplifier. The output of each amplifier is individually addressable. The space required for the converter and amplifier cause the light sensitive area to be proportionately smaller than found in a CCD array.*

In addition to the chip elements just described, CCDs have gating circuitry that shifts the buckets of charge, and there are areas between adjacent pixels that prevent a large amount of charge on one pixel from spilling onto near neighbors. This is called *anti-blooming*.

CMOS technology (see Figure 14.4) is the same sort of device architecture that is used in making computer chips. They rely more on voltages than currents. Each pixel is virtually the same as it would be if it were on a CCD, but each has its own ADC and there is circuitry that allows us to randomly access any pixel in the array. The anti-blooming feature is included. The data coming off the sensory portion of the chip is known immediately because the external circuitry has to probe the chip to get the data. The information is immediately in digital form and calculations on it can start right away.

COMPARISONS

CCDs operate by moving charge, and the movement of electrical charge is, by definition, electrical current. In other words, CCDs are current-based devices, which operate on the basis of voltages, Current times voltage is electrical power. By comparison to CMOS devices, CCDs consume significantly more power. Generally speaking, the batteries in a CMOS camera will last a lot longer than those in a CCD camera.

CCDs have a single ADC and there is only a minimum of circuitry on the image-sensing area, per se. CMOS cameras have a large amount of control and access circuitry attached to each pixel, and this takes up valuable space on the sensory portion of the sensor chip. The result is that CCD chips use a larger proportion of the area of the light-sensitive portion of the chip for pixels. This means that they tend to have lower noise and higher sensitivity. Also, with the CCD array, the adjacent sensitive areas of the pixels are closer to each other. This means that there are fewer gaps in the sampling of the image.

The image projected by the lens onto the imager chip comprises a continuous analog signal. The pixels each sample a portion of that image. The size of the pixel is one determinant of the fidelity of that sampling. Basically any part of the image that falls on the active face of a pixel is sampled once and it becomes the sample for that whole portion of the image. But the pixels do not exactly touch each other. There are gaps between adjacent pixels. This is like the grout between tiles. The image is not sampled in these gaps. When there are gaps in the sampling, it is called *undersampling*. When every possible portion is sampled, it is called *critical sampling*, and in some cases where areas are sampled more than once, that is called *oversampling*. Digital cameras undersample. CMOS chips undersample more than CCD chips because the spaces between the pixels are larger. The fact that each pixel can respond to only one of the three primary colors results in a color-dependent undersampling as well. The red and blue light is undersampled more than the green. The algorithms used to compute an image from the raw file must compensate for the undersampling.

It is important to note that with both CCD and CMOS designs, the sensitive areas of the pixels do not touch. There are gaps. Each sensitive area is covered by a filter and these are shaped like small lenses. This helps to funnel light that falls on the gaps into the sensitive areas. But even these do not touch. But they do have the effect of increasing the sensitivity. They also tend to increase the depth of field one will get with a given lens and lens opening. This is because some of the low blur associated with slightly out of focus items is made discreet.

With both types of chips, it is possible to change the voltages applied to the pixels and thereby alter their apparent sensitivity. Most cameras have a control by which the photographer can change the ISO sensitivity.

The nature of the fabrication processes for CMOS technology is generally cheaper than that for CCD technology. The result is that CMOS chips are a bit less expensive than CCD chips.

CMOS technology allows direct access to the digital data for each pixel. This means that the chip is capable of programmed control. For example, you might sample only a subset of the total pixels—for example, the 50% of them that are in the center of the chip. Then you could interpolate up to produce a full array of output pixels. This is called *electrical zoom*. It has intrinsically lower resolution than *optical zoom*, but it is somewhat easy to do in the camera. This would be harder to do with a CCD-based camera.

With CMOS technology, the ability to select pixels at will, or independently set the parameters for each pixel, allows for some very special characteristics. The sensitivity of a pixel can be altered by adjusting the control signals applied to it. In addition, it is possible to harvest the contents from a pixel very quickly and then have it continue to collect charge. These properties make it possible to dynamically alter how selected pixels respond. For example, you could program the chip to empty a pixel of charge when it reaches some level close to saturation. That initial emptying is noted and then the pixel is allowed to continue to record additional incoming light. This results in a sensor chip with a large dynamic range. It can operate all the way from the dark current limitation on the low light end to at least twice the saturation point. This approach is beginning to show up in certain specialized video cameras. In fact the pixels can all start out biased for high sensitivity and, after the saturation point is reached, the bias will be readjusted to a lower sensitivity. The result is a greatly expanded dynamic range. With digital still cameras, where it is desirable to have shutter speeds of 1/2000 a second, it is hard to implement the technique of emptying pixels mid-exposure.

OTHER APPROACHES

One manufacturer uses octagonal pixels instead of square ones. Those pixels are arrayed in a octagonal array as well, as shown in Figure 14.5. Finally, the chip has pixels with different sizes. The octagonal pixels arrayed on an octagonal grid mean that there is some degree of overlap across lines of detail from the image. In the case of a vertical image line, one row of pixels

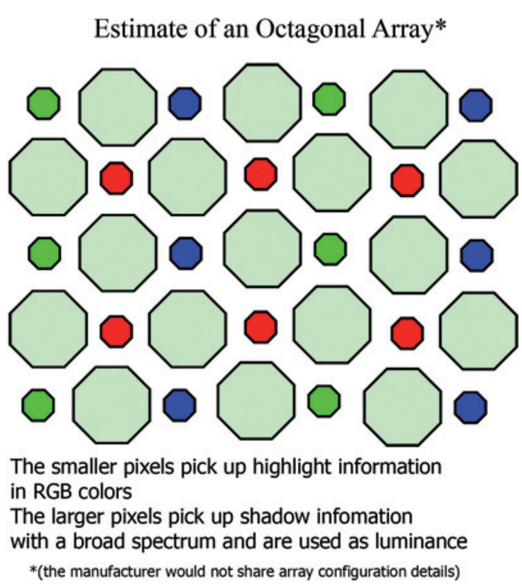


FIGURE 14.5 *Estimate of an Octagonal Array. Some digital cameras use pixels that have an octagonal shape. These chips also have equal numbers of large and small pixels. If the small pixels had specific color sensitivity, they can be used to estimate the colors needed for a color image. Since they are small, they would respond primarily to high brightness light. The larger pixels will respond to lower light levels, where color information is less important than luminosity information. These pixels could have broad color response.*

might align such that the line edge lands between adjacent pixels. In the octagonal array, the next row of pixels will be such that at least part of a pixel will straddle the line edge. With proper algorithms applied to the image data, the overlap reduces the effect of the problem of undersampling inherent in the square arrays, and images can be computed to have higher resolution than might be expected. In fact, the undersampling is still there, but the information from adjacent elements allows a better estimate of the missing information. These chips utilize CCD technology.

In addition to the undersampling issue, the octagonal array allows for some of the pixels to be larger than others. The larger pixels have higher sensitivity (due to their larger area) than the smaller ones. But they are all responding to the same image at the same time. The result is that there are two records—a high sensitivity one and a lower sensitivity one. The two records can then be combined such that the high sensitivity pixels control the formation of the low light level portions of the image, and the low sensitivity pixels are used more to compute the highlights portion. The result is a greater dynamic range.

There is another approach to designing sensor chips. In essence, the pixels sensitive to the three different primary colors are stacked one on top of the other. The depth to which a photon penetrates before it triggers the loosening of an electron depends on the wavelength of that particular photon. To implement this approach, the sensor chip must be able to determine the depth at which electron/hole pairs are formed. From this, the system can assign a color to that incident photon. This approach obviates the need to estimate the values for the other two colors, as with a Bayer array, and it also reduces the undersampling problem. The result is that we can achieve images with apparently greater sharpness. There is only one manufacturer using this technology today.

Scanners have sensor chips that have three or four rows of pixels, each in a long, single-file array. These pixels can have color filters on them like the Bayer array pixels, or they can rely on a spectral separation of the incoming light (see Figure 15.7). In this approach, there is no need to estimate the other two colors since all three colors are independently sampled for each location. These chips are much cheaper than the rectangular ones and they capture the image one line at a time. This is not a viable option for an object that might be moving and therefore cannot be used in a digital still camera. Some video cameras have three chips as well—one for each color—but they are much smaller than those used in digital still cameras and video has much lower resolution.

In summary, for all but two manufacturers, the basic difference is whether the camera uses CCD or CMOS technology. The only real difference the user is likely to see is that the batteries last longer with the CMOS cameras. Compensation for all the other differences largely makes it so the user is unaware of the sensor chip choice. The other option by a major manufacturer is the octagonal array cameras. They are highly competitive with the Bayer array cameras. The stacked sensor cameras are available from only one vendor and enjoy a very small market share at this time.

The output from any of the sensor chips is a string of numbers. Those numbers have information gaps, and the data must be processed to compensate for the gaps by making estimates of what the missing information should be. The compensation processing is part of a bigger process of analyzing the raw image data to obtain information that a display or a printer can convert into a viewable image. All the Bayer array and octagonal array cameras are capable of outputting industry-standard JPEG images. They all are capable of outputting raw files as well, but in this case the camera manufacturer's file converter, or a high-quality substitute, is required in order to get an industry standard image file such as a TIFF.

A NEW DESIGN

At the time of this writing (2007), the Eastman Kodak Company announced a new chip design that will probably show up in cameras about the time this book is released. This design will offer enhanced photographic sensitivity and increased dynamic range. In the Bayer array, there are pixels that are made sensitive to red, green, and blue light by placing filters atop the sensors themselves. The premise for this design is derived from human vision. Basically, people see outlines of luminance and then color in the images using color sensors in the retina. Scene composition is done in the brain, not in the eye. The eye has many more green and neutral sensors than red or blue sensors. The Bayer array has twice as many green sensors as red or blue. The image is compiled by outlining it largely with green light and then layering in the color information derived from the red and blue pixels.

The new chip design takes this a step further. The array is modified to add pixels with no attached filters. These are treated as if they were visual-neutral pixels. Remember that an overall infrared filter is used to ensure that long wavelength light, invisible to people, does not dilute the responses of the pixels. So the infrared is blocked. Also, silicon, the native material from which the chips are made, is not very sensitive to ultraviolet light. So even if some short wavelength light comes in, it does not do much. The result is that pixels with no integral filters will be close to being visual-neutral sensors. And since they have no overcoat filter, there is no filter factor. The result is that these pixels can be two to four times more sensitive than those with filters. The end result is that the visual-neutral filters have high sensitivity.

The nonfiltered pixels will be able to see detail in areas that are too dark for the filtered pixels. That is to say that they will pick up detail in shadow areas of the image. This is consistent with human vision as well. The more sensitive rods are used to see dark detail and do not resolve color. The nonfiltered pixels are then used as visual-neutral pixels, along with the green-filtered pixels, to determine the fine structure in the image. Finally, the red and blue pixel data is used to fill in the coloration of the image. This approach could be a significant step forward in digital photography.

There have been attempts at linking images to cameras for forensic purposes. The approach is to map the locations of both hot pixels and dead pixels in the suspect camera and then examine the images to see if they show corresponding bright and dark pixels. The results have had limited success. In order to obtain a viewable image from a camera, information from several pixels in the vicinity of each pixel has to be combined to estimate all three colors. This is the *demosaic* process. Combining the information from several pixels tends to reduce the visibility of the performance of any single pixel. If the image is rendered as a JPEG image, then it is even harder. Even if the image is taken as a RAW file, care must be taken to identify which pixel is which in terms of the Bayer array, and then you must compare red pixels to other reds, greens to greens, and blues to blues. Finally there is the matter of separating image detail from pixel performance. The end result is that the process is very difficult and the outcomes are mixed.

CHAPTER 15

Storage and Media

Images captured in a camera are stored by the camera on a solid-state memory device that can be removed from the camera and inserted into a computer. The images are stored on these devices for a short time so that images and associated data can be transferred to a computer for further work; the files contained are called *primary images*. Once the image files are brought into a computer, they may be recorded onto writable compact discs. These files normally serve the purpose of providing a portable but frozen version of the original files and are used by the examiner as he or she works the case. The files may also be put onto a secure server, marking the start of the *archiving* process. When a primary image is copied onto a medium that will be treated as permanent, the new image becomes the original and the primary image can be erased (this is described in Chapter 18, as well). Servers use magnetic hard drives and involve security and file-management capability. In summary, the camera-removable media are short-term devices, and the images should not be kept on these for more than a few weeks at most without backing them up. The CD resident files are intended for use while working the case and can be used for a few years. The long-term, archival records should probably be kept on a secure, managed server.

When debating the various media options, consider the following properties:

- **Instability**: An indication of how long the data can be reliably stored when the media is properly maintained.
- **Capacity**: An indication of how much data can be stored without compression.
- **Convenience to write and maintain**: An indication of how hard or easy it is to put data onto the media and retrieve it.
- **Ability to protect**: An indication of how hard or easy it is to change data that has been stored.

It should be noted that when a file is duplicated or copied using a typical operating system and computer, the system will automatically check to ensure that the data are transferred accurately. If an error occurs in the process of transferring a file, the computer will provide a warning that the process might be faulty. Such indications should always be taken seriously while the file to be transferred is still available.

REMOVABLE MEDIA

Digital Camera Flash Cards

In every computer there are at least three types of memory. We are all quite familiar with two of these: the hard drive and the RAM (random access memory). But there is a third type as well: the chip that holds the BIOS (basic input/output system). The hard drive will retain the data written to it when the power is shut down (nonvolatile) and the RAM will lose its data as soon as the power to it is gone (volatile). The CPU (central processor unit) looks to the hard drive and the RAM for the information it needs to operate, but when the machine is started up, it does not have the ability to read the hard drive and import the operating system—its basic instructions on how to operate. So to get things going, the BIOS device holds nonvolatile information and has the ability to get the CPU to be able to read the operating system from the hard drive. This process is called *booting up*. The BIOS device needs to have nonvolatile memory. The technology employed is called *flash memory* and is very much akin to that in memory cards that go into digital cameras, thumb drives, cell phones, and many other devices. Unlike the hard drive, it has no moving parts, is quite small, and does not use much power when operating. But, per megabyte, it is much more expensive than a hard drive, so hard drives continue to be used.

Flash card technology was invented by Dr. Fujio Masuoka in 1984. At the time, he was an employee of the Toshiba Corporation.

How Flash Memory Works

A flash memory consists of a rectangular array of pairs of field-effect transistors (FET) that are linked like Siamese twins. One of these transistors is connected to what is called a "word line" wire (think of it going across rows); and the other is connected to the "bit line" (think of this as going along columns). A FET has three component parts: a source, a drain, and a gate. Operationally, current flows from the source, past the gate, and on to the drain. The amount of this current is controlled by the voltage on the gate. In a flash memory device, the gates of the two transistors are connected by a metal oxide layer (similar to a very thin glass insulator). The transistor connected to the bit line is called a *floating* gate because its gate is connected to the word line only through the other transistor, called the *control gate transistor*, which is directly connected to the word line.

In its initial state, there is no barrier to stop one transistor from affecting the other. This state is given the binary value of 1. When a sufficient voltage is applied to activate the floating gate transistor, its gate acts like an electron generator. The design of the device is such that it allows tunneling. *Tunneling* is a quantum mechanical process by which electrons can, in very special circumstances, penetrate an insulator. The result is that when the bit line is activated, electrons are generated in the floating gate transistor, penetrate the metal oxide barrier, and build up on its other side. This build-up prevents the control gate from acting. A cell sensor monitors how much charge is built up, and when it is above a certain level, it is a binary 1; if it is below that level, it is a binary 0. Since the special circumstances required for tunneling to occur are not present once the charge is deposited, it will sit on the barrier until it is removed in an erase mode. This is done by action of the word line, and at least a whole word is erased at once. The result is that flash memory will hold its data after the power is removed.

One key limitation of flash memory is that each chip has a finite life. It is capable of a certain number of read/write cycles and then it becomes unreliable. The number of cycles is fairly large—about 100,000. This is based on the number of cycles to a particular block on the chip. Using a controller that dynamically selects blocks based on how often they have been used tends to make that actual useful life appear much longer. Experience and theoretical estimations indicate that the chip, in practical use, will outlast the useful life of the camera. But flash memory should not be used for purposes other than short-term storage of critical data.

Flash Memory Properties

There are several versions of flash memory and, by now, several generations. But three properties are important to keep in mind. We can write to the memory device one bit at a time, but erasure must be done to a whole block at once. Second, the array holds a vast number of elements and in order to keep costs low, some of the elements are faulty. In operation there must be intelligence that avoids using the blocks with faulty elements. Finally, since the conversion from a 1 to a 0 is based on passing a certain threshold level, there is room for error. To ensure that data is not contaminated, the system needs EDAC (error detection and correction). With these factors taken into account, flash memory is a reliable, nonvolatile computer memory that can be electrically erased and reprogrammed.

The popularity of flash memory can be explained by its fast read (access times) and its shock resistance and relatively low power consumption. These features make it ideally suited for mobile devices, such as cameras, audio players, and mobile phones. Another excellent feature of flash memory is that when packaged in a memory card, it is quite durable—able to withstand a wide range of pressures, temperatures, and even immersion in water. Exposure to strong and rapidly changing magnetic or electric fields could cause some data loss, but this is a rare situation.

There are several other reasons for using flash memory. It is small in size, light, noiseless, and has no moving parts. Also, flash memory supports plug-and-play on compatible computer operating systems; so it can be inserted into a computer or a card reader and be recognized and accessed quickly by the computer. They also support hot-swapping, which lets you plug or unplug devices without needing to shut off power and then restart the computer. This feature enhances the portability and convenience of flash storage devices for transferring data, pictures, or music between two computers or devices. (It should be noted that when flash memory is packaged as a USB thumb drive, it cannot just be pulled from the computer—the computer's software must be set to release the device.)

Flash File Systems

Flash memory is written and erased in different ways. Writing is done bit by bit via the bit line, and erasure happens a block at a time via the word line. Also, to avoid errors, there must be a formatting process that locates and isolates defective areas of the chip, and there must be error correction and detection software and redundant information. The result is that special file systems and data controllers are employed. The cards specifically designed to work with digital cameras, such as SD and CF cards, have a built-in controller. Others will need a specific device driver in the computer working with the device.

To write to a disk, the computer must temporarily copy data from the chip, update the information, erase the whole block, and finally, record the new data. This also requires revising the file pointers and table of contents. Newer flash card devices use the FAT (file allocation table) process, which makes the device look like a hard drive to the host computer.

Flash memory devices must be formatted prior to use. Formatting includes the following operations:

1. Each memory cell in the device is tested to ensure that it is working properly.
2. All defective cells must be identified and provisions made so that they will not be involved in any data transactions.
3. Some cells are set aside as spares. These are held in reserve to replace any memory cells that may fail over time.
4. A directory must be created, usually a FAT, which also is used on hard drives.
5. Some cells are used by the flash storage device's controller for storing specific information used by the controller.
6. Depending on the type of flash card, some cells may be reserved for special features such as copy protection.

When flash memory is used in a computer, it will be formatted by the computer. When the disks will be written by a digital camera, it should be formatted by that particular camera. This maximizes the utilization of the device's capacity.

Capacity

The specific elements in a memory card may be composed of smaller units, which can range widely in their capacity. To achieve higher capacities, several smaller chips might be combined using the same processes used to manufacture other integrated circuits.

Speed/Performance

Flash memory cards are available in different speeds. Some are specified with the approximate transfer rate of the card, such as 2 MB per second, 12 MB per second, and so on. Others might be rated in terms of a multiple of the rate at which a CD transfers music—150 kilobytes per second. These might have ratings of 100X, or 200X, and so on.

Flash card storage device performance usually depends on the following three factors:

1. **The specific flash memory chips used**. Usually, there is a tradeoff between the high-speed and more expensive single-level cell (SLC) flash chips, and the standard speed and more affordable multilevel cell (MLC) chips. Check the specifications on the card prior to purchase.
2. **The flash storage device's controller**. Most flash storage devices have a built-in memory controller, which manages the interface to the host device. If the host controller is capable of supporting faster data-transfer speeds, it can result in significant time savings when reading or writing data.
3. **The host device to which the flash storage device is connected**. The host device can limit the specific read and write speeds. Therefore, using faster flash storage devices will not increase performance. An example would be if your computer supports only the slower USB speeds, a hi-speed USB drive will not result in faster transfers. Computers must be properly configured to support faster transfers in both hardware and software.

Digital Camera Media Cards

Although there are too many types of flash cards (at last count, 15) to cover here, we will mention a few of the more common cards that are available for digital cameras.

Compact Flash Cards (CF)

CF is the oldest and most successful type of card. It presently holds onto a niche in the professional camera market. It has benefited from having a good cost-to-memory-size ratio and also from having larger available capacities than smaller formats.

There are two main types of CF cards: Types 1 and 2. They differ only in physical size; the Type 2 card is thicker. There are four main speeds of cards, including the original CF, CF High Speed (CF2.0), a faster CF 3.0, and a yet-faster CF 4.0 standard that was introduced recently.

A CF card can be used in a computer in a variety of ways. It might be used directly in a PC Card slot when inserted into a plug adapter. Or it can be used as an IDE hard drive with a passive adapter. Finally it can be used in readers that will connect to any number of common ports such as USB or FireWire.

Compact flash cards are generally available in capacities from about 512 MB to about 64 GB. The most popular choices seem to be between 512 MB and 8 GB. Lower-capacity cards, those below 512 MB, are becoming rare in stores, as higher-capacity cards are readily available at essentially the same price.

CF cards also are considered far more rugged and durable than many of the others because they have their contacts hidden, are relatively thick, and—being a bit larger than other types—are easy to handle. They are known to deal easily with shocks, impacts, and accidents. Recently a compact flash card was found to be still working after spending two years in a camera submerged in shallow water.

As with any device, there are some weaknesses, such as:

- Compact flash cards lack any mechanical write-protection switches.
- Improper insertion can cause damage to the receptor device, but this rarely happens since the slots are designed to prevent incorrect insertion.
- The large dimensions, in comparison to other cards, limit their use in very slim devices.

SmartMedia

The technical name for this is actually solid-state floppy-disk card (SSFDC). SmartMedia is very small—37 mm wide by 45 mm long by about 0.75 mm thick. It is easily the thinnest of the flash memory formats. It has a notch in one corner and exposed gold contacts on the back side. The SmartMedia cards can be read in most card readers and PC slots with the use of an adapter.

A disadvantage of SmartMedia is its storage capacity. The maximum capacity you can expect to find for SmartMedia is a mere 128 MB, making it a much less appealing solution for modern cameras storing 10 MB image files. Another shortcoming is the exposed contacts, which can become dirty or worn.

xD Picture Card

This format was intended solely for use with digital cameras, although it did find some additional applications. xD cards are smaller than most of the other cards, at 20 mm by 25 mm by 1.7 mm. The maximum capacity is expected to be 8 GB, but typical card capacities are in the range of 1 GB in size. The read speeds are up to 5 MB/s, and write speeds can be up to 3 MB/s, making them fast, but not the fastest. xD cards usually are promoted not only for their minimal size, but for their low power consumption.

Memory Stick

This flash memory is used in Sony brand devices, including digital audio devices, cameras, and even televisions. Memory stick flash cards are a little smaller than a stick of gum, and measure approximately 50 mm by 21.5 mm by 2.8 mm. The capacities range from 2 GB to 8 GB. Theoretically the maximum data-transfer rate is 160 Mbps, but real-world results will probably be lower. They are used in video devices as well as still cameras.

Memory sticks come in four types: the original memory stick, memory stick PRO, and Duo versions of each. Memory stick PRO offers faster speeds and larger capacities than the original memory stick. The Duo modules are smaller and actually use an adapter to fit into memory stick slots. Note that not all devices that take memory sticks can use memory stick PRO modules—be sure to check your manual.

Multimedia Cards (MMC)

Multimedia cards are some of the smallest flash cards available, about the size of a postage stamp. They were introduced initially to be used in the mobile phone and pager markets. Now they are commonly used in digital cameras. The newer version of MMC cards are backward-compatible with the older MMC cards. MMCplus and MMCmobile cards offer higher performance than older MMC cards, and MMCmobile cards support lower voltage applications such as cell phones. MMC cards are designed for use in many popular portable devices. Their physical size has made them very popular for cell phones and other small devices.

Secure Digital Cards (SD)

Secure Digital is a second-generation derivative of the MMC standard. It has several important technological advancements compared to MMC. In particular it has the ability to use cryptographic security protection for copyrighted data/music.

SD cards are slightly thicker than the original MMC cards. This means that devices designed to support SD cards may also accept MMC cards.

However, devices exclusively designed for MMC cards will not support the thicker SD cards at this time.

Bottom Line

There is no standard among manufacturers for media storage devices in cameras. A few have been described here but, as was mentioned earlier, there are at least 15 different types today and no doubt there will be additional cards soon. If you have a fleet of devices that requires you to have several of these different types of flash memory, interfacing with all of them could be inconvenient since each has a different size and shape, and requires a specific slot. A good solution is a card reader, such as a 15-in-1 flash memory card reader.

This, with the help of some adapters, can read all flash memory. Most of the multiple-card readers are USB compatible and they are compact devices that require very little space.

Summary

Flash card memory devices have become reliable, have large capacity, and are fast and easy to use. They are not designed for long-term storage and it is considered unwise to rely on them alone to keep important data. The expectation is that they will be used to take images from the camera to a computer or other device to make a more permanent copy of the information.

CDs AND DVDs

History

The Compact Disc, or CD evolved from an idea in the mind of James Russell—working at the Battelle Memorial Institute in Richmond, Washington in the early 1960s—to its first patent in 1970, and has become a worldwide phenomenon today. In the early 1980s, as the result of collaboration between Sony and Philips music recordings were published on CDs. These were based on a prerecorded disc and a basic player that could be connected to a home music system. The information was recorded in a format called CD–DA (compact disc–digital audio), which is still in use today. Later that decade the CD–ROM (compact disc–read only memory) was introduced for publishing data on CDs. This technology is also still in use today. The physical aspects of CDs are standardized under ECMA-119 and the logical format is covered by ISO 9660, still in effect.

The next major advance was made in the introduction of the CD-I (CD-interactive) and later the CD-ROM/XA (CD–read only memory extended architecture). This was the beginning of the user being able to record the disks at his or her workstation. Not long afterward came CDs that could be both written and erased at the desktop. For reasons of security and permanence, the rewritable disks are not of interest for forensic use and will not be discussed.

Early in 1990, collaboration between Kodak and Philips resulted in the introduction of the Photo CD. This was the first time that the amateur photographer was invited to have photos on a CD that could be played on a home television or computer. Kodak sold desktop writers, CDs, and software for both the home and professional photographer. Unfortunately there were not that many home computers at the time, and Kodak was prohibitively protective of the formats (preventing widespread standardization), and so the product never gained sustainable market penetration. Nonetheless, the idea of putting photos on a CD at the desktop was born.

Kodak was acutely aware that people do not want their photos to fade, and that professional photographers will want to be able to protect their images. Accordingly they developed discs with special long-life characteristics. The deformable dye layer in the CDs were made of a special polymer whose evaluations indicated that the material would be stable for about 100 years under nominal keeping conditions. They also used gold, an ultra-stable metal, instead of aluminum, a highly reactive metal. To deal with making the images identifiable, each disk had a machine-readable bar code etched into its gold layer. These properties were also attractive to law enforcement applications, and in 1994, Kodak introduced the QuickSolve system, specifically designed for law enforcement applications. The system was developed in Indianapolis but sold primarily to federal agencies. In addition to QuickSolve, Kodak was selling SLR-format digital cameras to forensic investigators during this period. The use of digital imaging in forensic applications has grown dramatically ever since (although Kodak is no longer in this specific market). The special, ultra-stable CDs are also no longer available, and users must be wary of the CDs they purchase. Since the longevity of CDs is no longer known, they are not recommended for long-term archiving of images. A few years' time is fine, but a few decades would be very risky.

Digital Versatile Disks, or DVDs, are based on technology very similar to that used in CDs. They use newer technology and can hold about four times as much information.

Basic Technology

Since we are interested primarily in CDs that can be written by the investigator at his or her desk, we will not delve very far into prerecorded CDs. Likewise we are not interested in rewritable CDs. The focus is on CD-Recordable (CD-R) technology and its counterpart in the world of DVDs. Since these devices can be written once and read many times, they are dubbed WORM devices (write once read many times).

The Data Rates Associated with Disk Reader/Writers

The common way to notate these rates is as a number followed by an X. So a drive might say that it is a 12X device. This is read as "twelve ex," and

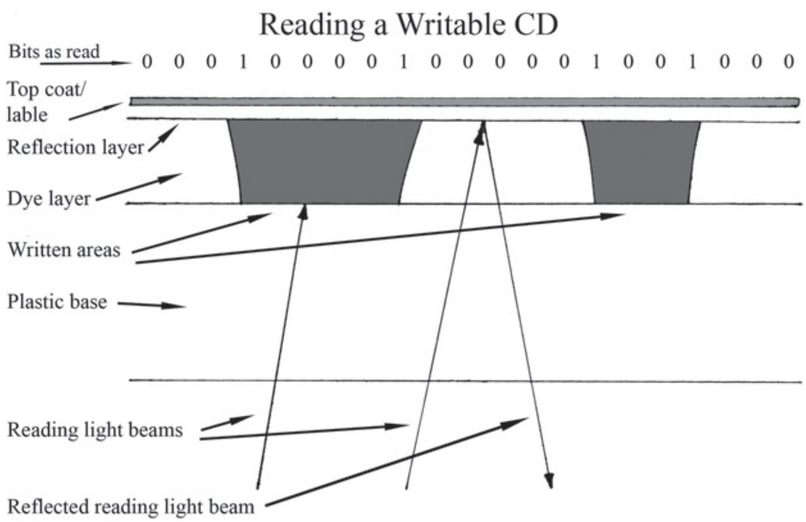


FIGURE 15.1 *Reading and Writing a Writable CD or DVD. The assembly of devices shown is in reality quite small and mounted on an arm so that it can move in from the circumference of the disk or out from near the center. The disk is spinning and in the diagram the area being written is moving out of the page. The read right head will move left and right.*

means that the data flow rate is 12 times as fast as the rate that would be used in playing music from a CD or playing a movie from a DVD. In other words, since CDs were developed for music and DVDs were developed for movies via video, these define the respective base data rates. In the case of images, which are data files, the faster the data rate, the better.

The typical CD-R has four active layers, as shown in Figure 15.1. First of all, there is a clear plastic (usually polycarbonate) disk that supports all the other layers. On top of this there is a special polymer layer, which is the layer that will change in response to the writing action. Above this there is a reflective layer. Finally, there is a protective layer on top. Labeling may be placed on top of the protective layer.

To write to the disc, it is first made to spin. As it does, a laser is focused through a complex set of optical elements onto the polymer layer. In the most common devices, the layer is transparent to start with and turns opaque when the laser is on. The darkened material is not removable. To read the disc, the disc must be spinning, and a laser of much lower intensity is focused on the polymer layer. The write laser has to physically alter the polymer, and the read laser must have a power level that prevents it from altering the polymer layer. Where it has not been darkened, the light goes through the layer and reflects back off the reflective layer. In areas that were darkened, the light does not reflect back. A photo sensor is monitoring the reflected light. The dark areas are referred to as "pits" and the other areas are called "lands" (after the nomenclature developed for the original prerecorded CDs). A transition from a pit to a land or from a land to a pit is assigned a value of a binary 1. The lack of a transition is a binary 0.

Since the user of the disc-writer chooses the discs to use, the device cannot always correctly set levels to write the data. If the laser has too much power, the polymer might be destroyed, and if it is set too low, it will not write effectively. In order to ensure proper writing each time, the drive will write a short test track ahead of the table of contents (the first entry on the track). It will adjust the laser power to achieve reliable data-writing.

The pits are recorded in a track that starts at a point close to the inside diameter of the recordable area. The track then follows a spiral path outward to the outer end of the recordable area. Both reading and writing are accomplished by following this spiral. At small radial distances from the center of the disc, the disc spins at a relatively high speed, about 500 rpm. As the site of reading or writing moves out from the center, the rotational speed decreases to about 200 rpm. The purpose is to keep the linear speed the same all the way along the radial path. This change in rpm with location along the disc radius is referred to as the *constant linear velocity* (CLV) approach; this was the way all drives and writers were built at first. As the data transfer rates were increased with newer technology, it became better to use the *constant angular velocity* (CAV) approach, which is used in hard drives. This means that the drive has to have means to correct the read bit stream and convert it to the same result we would have if using the original CLV standard. The same discs can be used in either type of drive and the only thing that the user might want to know is that if the drive is operating at 12X rate or more, the device is probably using CAV technology.

The length of the pits will be varied in accordance with the data being recorded. Likewise the length of the lands is varied. There is a clock in the reader/writer and since the linear speed of the track is constant, the ticks of the clock indicate equal advances along the track. Since the transitions indicate a binary 1, and these must be followed by nontransitions or 0s, the system can count the number of ticks of the clock between 1 s to determine the number of 0s; this is how the recorded value is determined. The information is in the lengths of the lands and pits. Conversion of the raw data from the disc to more traditional sequences of binary numbers is very complex and follows a protocol known as EFM (eight-to-fourteen modulation).

This basic principle is quite straightforward, but the implementation is quite complex. This stems from the incredible dimensions of the lands, pits, and tracks. The lands and pits (and, as a result, the track) are 0.5 microns wide. The spacing between the successive tracks is 1.6 microns. The pits and lands have a minimum length of 0.83 microns and a maximum of 3.3 microns. In other words, the specific information is recorded on a very small scale, yet a huge amount of data is recorded. If the track were laid out in a straight line instead of wrapped in a tight spiral, it would be half a micron wide and

5 kilometers long. Instead it is spiraled onto a disk 12 centimeters in diameter with a 1.2 centimeter hole in the center. There are about 20,000 windings of the spiral. The usable packing density is about 1 megabyte per square millimeter.

The geometry of the disks leads to the complexity of the technology needed to reliably read and write data. The most important issue is keeping on the track. As the disc is being written or read, the device must be sure to stay on the proper track. In general this is accomplished by having multiple light beams. One (or more) is reading the center of the track and the others are reading just outside the edges of the track. By keeping the edge beams such that they do not read data while the central beam does, the system stays on track. When writing, the outside beams must not pick up data from the adjacent, inside track. The data to be written must be fully assembled and processed before it is written. This allows the software to break the data stream into blocks.

Since errors in the raw data are altogether possible, precautions are taken in writing data CDs to prevent the errors corrupting the information. One precaution that is taken is to spread the data into locations spaced relatively widely. A string of numbers, for example, will be broken into parts and the parts are located at various noncontiguous locations in the track. This is called *interleaving*. In this way, a scratch or particle of dirt on the surface will not obliterate a whole block of raw readings. In addition EDAC (error detection and correction) techniques are employed. The exact methods are extremely complex, but the idea is that there is a certain amount of a pattern of redundancy of recording, such that if some raw readings are lost for any reason, the initial stream of data can be reconstructed. The addition of data redundancy also requires assembling all the data prior to writing it. Music CDs do not suffer as much from errors as data CDs, so there is less redundancy on music CDs, and the total amount of stored, usable data is higher. Each data block on a music CD can hold 2352 bytes, whereas a data CD block can hold only 2048 bytes.

Modern drives use complex optics to read more than a single spot at a time. This gives effectively a higher data-transfer rate. The light from the laser goes through a diffraction grating that splits it into as many as seven separate beams. The beams are then sent through a beam-splitter mirror. The energy that passes through goes through a collimator lens and possibly a polarizer, which rotates the angle of the light going on through to the polymer layer. Some of the beams are used to read one spot, others are used to ensure that the optical system stays on track, and still other beams can read other spots. The reflected light is rotated again by the polarizer, and reflected off the beam-splitter mirror to the array of photo detectors. Buffers in the reader/writer ensure that the data stream is continuous as it leaves the device.

Writable DVDs come in two basic styles: DVD–R and DVD+R. The choice depends on the specific writer the user has. Most of the newer drives will take either disc. The resulting discs can be read in any modern reader.

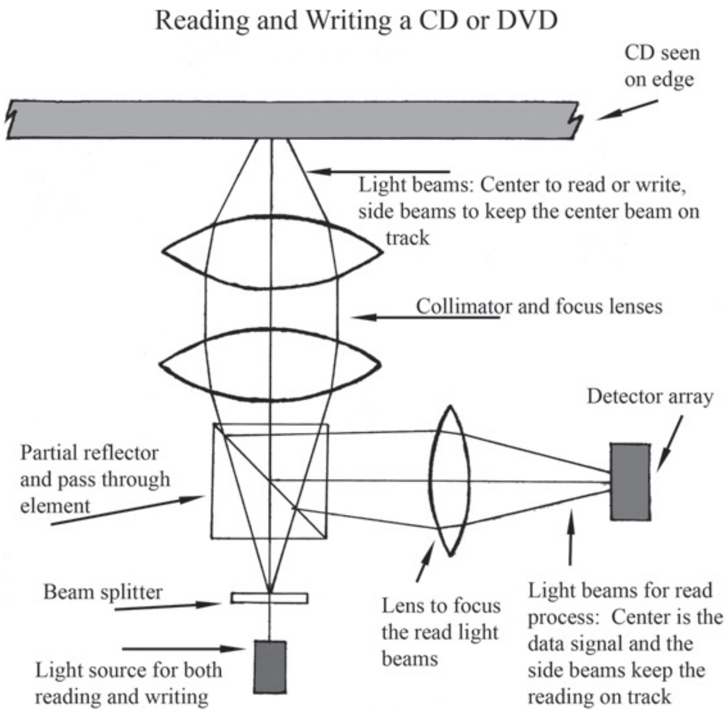


FIGURE 15.2 *Reading a Writable CD. The read beams are shown entering from the bottom, through the plastic support for the disc. In areas where the writing process had made the dye layer dark (burned), the read beam is absorbed. In areas where the dye layer is transparent, the read beam goes through, bounces off the reflection layer, and comes back out to the read head. The areas where the reflection changes are read as 1s and other locations are read as 0s. The number of 0s between 1s carries the data.*

The other factor has to do with the amount of data that can be put onto a disc. There are single-sided, single- and double-layer discs, which hold 4.38 and 7.95 GB, respectively. In addition there are also double-sided, single- and double-layer disks, which hold 8.75 and 15.9 GB, respectively. These compare to a data CD, which holds about 0.7 GB.

DVDs were developed specifically to hold movies, and their parameters are designed specifically to be compatible with video technology. However, they can be used for both music and data as well.

The main differences between the basic technology in DVDs and CDs are that the pits and lands are smaller, the tracks are closer together, and, as a result, the tracks can be made longer and hold more data. The single-sided, single-layer DVD is constructed in essentially the same way as the CD. The single-sided, double-layer DVD has two different reflective layers, each with its own polymer layer. The first reflective layer the light beam encounters allows some light to penetrate to the lower layer. The other layer has a different reflector and this reflects back through the top layer. The first track spirals out from the center as in a CD, whereas the second layer spirals inward. This is to ensure that there is not a gap in the action in the movie.

Recently, to accommodate high-definition TV, two high-definition DVD (HD-DVD) formats were competing in the marketplace. One is called Blu-ray Disc and the other is Advanced Optical Disk (HD-DVD). Both rely on smaller features and a shorter wavelength laser to achieve higher capacity. Blu-ray has emerged as the winner, since the group sponsoring HD-DVD format has stopped new production. The name Blu-ray is based on the use of a writing laser that has a wavelength deep in the blue, around 400 nm, while earlier devices use a red laser that is in the range of 650 or 780 nm.

Summary

Relative to the performance properties set out earlier, we can say the following about CDs and DVDs relative to forensic applications:

- **Instability**. Assuming good-quality discs and writers, we can say that the data on these discs is stable for up to 10 years. Beyond this, the stability is subject to question due to the fact that the ultra-stable discs originally intended for photos are no longer being made. Several practitioners have had data on CDs for 10 years or more and no data has been lost.
- **Capacity**. CDs hold about 700 MB, and so about 100 photos with 7 MB files can be stored on a single disc. If the images are being stored in TIFF format, and if the camera used had 8 megapixels, the image files will be approximately 24 MB each, in which case each disc will hold only about 28 images. A DVD can hold at least four times as much data.
- **Convenience**. Good-quality read-and-write devices for both CDs and DVDs are now commonplace. Good-quality discs are also readily available, although you should be careful about purchasing bargain-priced, unfamiliar brands of discs. The computers treat the disc reader/writer as if it were another drive. There is software and processes that will allow the operator to close the disk so that no additional information can be added afterward. Some discs still have serial numbers and this can help ensure that the images are from the correct disc. Storing large numbers of discs requires advanced planning and a reasonable physical and logical file system.
- **Ability to protect**. Data written to a WORM, serial-numbered disc which is then closed, is quite secure. The files can be damaged, but they cannot be altered.

In summary, CDs and DVDs can be a good medium for storage of images while the case is active. It would be good practice to back up the files on a secure server, but the investigator can work with the discs conveniently and reliably.

NONREMOVABLE MEDIA

Hard Disk Drives

Hard disk drives offer magnetic, nonvolatile storage, and were first introduced in the mid-1950s. The hard disk drive (HDD) is both a hard disk and a hard drive, all packaged together in a sealed compartment. Unlike floppy drives and floppy diskettes, which are separate units, the HDD is bundled together and the user never separates the media from the drive. Therefore, the HDD has come to be called either a hard disk or hard drive, interchangeably. The hard disk is made of a rigid material coated in magnetic particles, typically iron oxides, that can hold a charge in concentric circles around the disk, called a *platter*. One magnetic charge or absence of a charge is stored in precise locations along the circular tracks. The positive charge (on) is a 1 and the absence of a charge (off) is a 0—called a *binary bit*.

The hard disk is made up of one or more platters (think of a phonograph record) stacked on a spindle inside the hard drive compartment. Each platter has two sides and its own read/write head, which never touches the platter (when it does touch, the result is a head crash—not a good thing). The heads are aligned together on an actuator arm that moves together along a path (in or out) by the disk controller. The data is stored in a series of charges (or lack thereof) around each concentric circle, called a *track*, and tracks are further divided into *sectors*. Each sector holds 512 bytes of data. A view of the geometry of a hard disk, showing the tracks, sectors, and cylinders is shown in Figure 15.3.

The platters and all the hard disk mechanisms are sealed into an airtight unit. The spindle spins the platters at a high rate (typically 5400 to 7200 rpm) with the heads floating just a whisper above the platters.

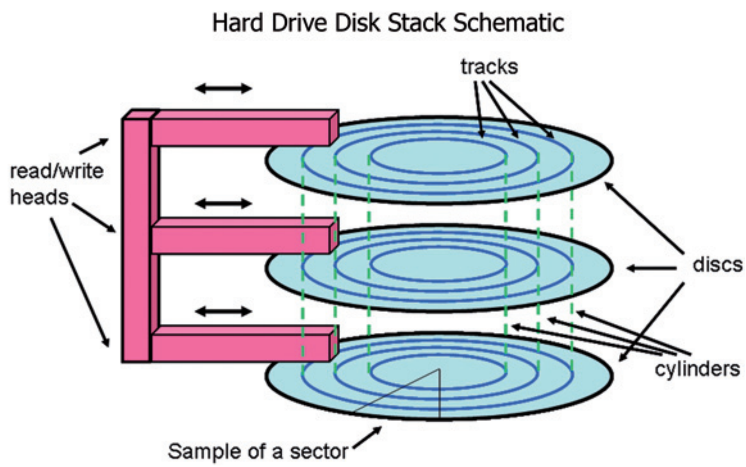


FIGURE 15.3 *Schematic of a Hard Drive Stack. The hard discs are shown in blue and the read/write assembly moves radially. Each disk has circular tracks and pie-shaped sectors. The corresponding tracks from all the dics form a cylinder.*

The unit is sealed because even a particle as small as a smoke particle can cause a head crash with resulting data loss. That is why a HDD should never be opened except in a "clean room" environment (free of dust and as clean as a surgical room in a hospital) and by skilled technicians—that is, if you want to keep your data intact.

A single disk can be further partitioned into additional drives using partitioning software. One drive can be partitioned into multiple drives, giving each drive a different drive letter, such as *c:drive, d:drive, e:drive*, and so on. Today there isn't the need to partition a large drive into multiple drives, since current operating systems and BIOS aren't as limiting in their ability to read large disk drives. Some of the older operating systems had limitations on the size of the drive that they could read without special software and partitioning. Drive size barriers ranged from 504 MB upward to 2 GB. Today the next limit may be imposed by the 64-bit and 32-bit operating systems. The 32-bit operating systems limit the ability to address greater than 2.2 terabytes and the 64-bit operating systems can address upward to 16 terabytes and possibly more.

HDD are not meant to keep your data indefinitely or permanently. Disk media does have a limited life—it won't last forever. HDD's MTBF (mean time between failures) might give an indication of the failure rate or life expectancy, but it is very hard to determine those numbers based on that statistic. Plus, the life of a drive may depend on many other factors, including how long it is left running continuously; environmental factors (dirt particles, heat, and humidity); the number of times the disk drive is started or stopped; and other relevant factors. Keep this in mind when deciding how to maintain your data's integrity and how often you plan to back up your data.

Interfaces

HDDs are connected to a computer by their interface—ATA, IDE, EIDE, SCSI, SATA, and so on. The interface will dictate the type and speed of communication between the HDD and the computer. Some interfaces offer faster communication between the drive and the controller, although the speed differences may not be noticeable to the user. However, the interface can be a source of failure in a network set-up, which would render data inaccessible until that failure is rectified.

Partitioning and Formatting

Before a hard disk can be used, it must be partitioned and formatted. The manufacturer of the hard disk has low-level formatted the drive before shipping the unit. This low-level format defines the drive's geometry and divides the blank surface of the disk into tracks, sectors, and cylinders. As explained earlier, the disk must be divided either into one or more partitions—carving up the one hard disk into logical drive letters. This partitioning allows a single disk to appear to be multiple drives. High-level formatting (normally what the user performs) prepares a logical drive for use by a specific operating system and sets up the file system on the drive for that system.

Hard drive capacities have grown significantly over the last few years. IBM introduced the 10 MB hard drive, which was thought to be too large and would never be used. Now, however, we can purchase 500 GB (1/2 terabyte) and larger drives for home use, with terabyte drives for business being very common.

Ever since the HDD was introduced, technologists have been perfecting and working on the reliability and performance of the hard disk drive technology. Businesses have been demanding better performance and reliability, and RAID is one of today's methods to achieve that goal.

RAID

Redundant Array of Independent (or Inexpensive) Disks (RAID) is a design of connecting two or more disk drives in combination with hardware and/or software to increase performance and fault tolerance. RAID can be implemented at a hardware or software level and there are many levels or *flavors* of RAID from multiple vendors.

Fault tolerance is a system designed for quick recovery of a software or hardware failure. When hardware (hard drive or hardware part) or software fails unexpectedly, a system with fault tolerance built in will respond to those failures in such a way that the system keeps working without any downtime (system stops working.) In reality, it is virtually impossible to create a system that will never fail and with "perfect" fault tolerance. However, many system vendors have created fault tolerance systems that cover many of the common system failures, enabling a system's continued uptime and allowing users access to their data even during certain hardware or software failures.

Many fault tolerant systems use one of the levels of RAID (discussed next) to offer protection through mirroring, striping, or other RAID technology for hard disk failures, plus backup systems and software implementation to cover multiple failure issues. Mirroring involves the exact duplication of the contents of one disk onto another. Striping involves separating the bytes in a data string into several smaller strings. The smaller strings can be saved in different physical locations. Pointers are created and kept to reconnect the small strings. This way, a whole string of data is not subject to a physical problem in the recording medium. EDAC (error detection and correction) generally is used to ensure that the original data string can be reconstructed.

Some of the more common levels of RAID are:

- **RAID 0**—striping, a simple data striped over several disks but no redundancy. There is higher performance but no fault tolerance and less data protection since, if a drive fails, all data in the array is lost.
- **RAID 1**—data mirroring, which gives two identical copies of the data on two separate physical disks. Offers high performance, is simple to implement, but expensive with twice as many drives required.

- **RAID 2**—this RAID is more in theory than in practice as it stripes data at the bit level and rarely is used since it would be extremely expensive to implement.
- **RAID 3**—a dedicated parity disk with byte-level striping. Used for very large data implementations and is less expensive than RAID 1.
- **RAID 4**—data written in blocks and parity information written to a dedicated disk. A common implementation of RAID. Parity data can be used to create a replacement disk if a drive fails.
- **RAID 5**—one of the most common or popular RAID implementations. Data is written in blocks onto disks and parity is generated and rotated around all disks in the array.

There are many more levels of RAID from multiple manufacturers. EMC and SUN have their own RAID levels for specific hardware configurations. The goal is to create a solid fault tolerance system for data integrity and storage.

Secured versus Reliable Data Storage

Secured storage is different from reliable storage. A secured storage system is designed to prevent or minimize unauthorized access to the secured area of a network, which could allow data to be compromised (stolen or altered). The secured network is created through the implementation of a firewall—a system of software and hardware to control and monitor network and Internet access to the network systems—and access controls to the physical hardware systems room(s). There are many different physical access control devices, such as access control cards, biometrics, key locks, digital locks, and any other type of access control to the physical room housing the network systems. Additionally, there are many types of software and hardware systems available that create a multilayered system to restrict access to the network and monitor its activity.

Reliable storage is dependent on the reliability of the system software, hard disk drive, and network. Is the hard disk drive or system likely to fail or cause loss or damage to the data? A firewall or secured storage is not required to implement a reliable storage system. If data is lost or damaged due to a software or hardware failure of the storage system, then the integrity of the data is jeopardized and no amount of security via a firewall can save the data. A RAID system (discussed previously) is designed for reliability of the data to reduce the risk of data loss due to hard drive failures. Secured systems refer to the strategy and design of security systems, such as firewalls, access control, and authentication systems.

The reliability of a system or hard drive depends on many factors, from environmental to physical. What types of physical access controls are in place to prevent unauthorized access to the physical location of the systems?

Is there a full firewall design implemented to prevent unauthorized access via the network or Internet? Are there automated temperature and moisture-level controls to prevent overheating or excess moisture which can damage hardware? What type of a backup system is in place for each type of data on the system? Are the backups incremental or full and is there only one copy? How often does the backup occur and where/how are the backup data stored? Has the backup system been tested to verify restores of the backup work correctly? The architecture, system design, and structure also play a major role in the reliability and security of the system.

The system design and network architecture dictate if and how fault tolerance is implemented effectively. With all the factors that must be considered for secured and safe storage of digital data, strong system management (highly skilled technology staff) is needed to manage all aspects of secured data. It is the technology staff, network administrators, database administrators, and all the other information technology (IT) staff that is responsible for creating and maintaining a reliable, stable, and secure network/data environment. There must be solid policy and procedures in place for routine backups of data, disaster recovery, security and firewalls, and any other relevant issues surrounding the network and data systems.

If you have highly valuable data, you may want to back up your data often and keep a separate backup apart from the IT department's routine backup. Keep in mind, however, that without proper authentication, your backup copy may not be considered an exact duplicate of what was lost. Hashing the file and documenting the hash may assist in proving the backup is an exact duplicate, depending on the nature of the data.

The following are guidelines to consider when using hard disk drives and securing data:

1. Back up your data often, before making any changes to software or hardware. Keep your backup copy in a secure and safe location.
2. Use anti-static procedures when working with your hard drive and computer. Keep an anti-static mat under your chair.
3. Keep magnets and magnetic tools away from your PC.
4. Use anti-static bags for your media, ground yourself when working with the hardware, and follow the manufacturer's recommended anti-static procedures or guidelines.
5. Use a firewall, anti-virus software, and anti-spyware software to improve the reliability of your computer.
6. Keep your hard disk drive stable—try not to bump into it or cause it trauma.
7. Make sure your room temperature and humidity are maintained at the recommended manufacturer's guidelines for your computer and HDD.

In general, high temperatures are much more risky for the hard drive than any other computer component.

8 For your computer and software, incorporate strong secure passwords and change them often for increased security.

9 Encrypt sensitive files.

10 Keep your software up to date with recommended patches and fixes from the manufacturer.

11 Don't write down your passwords on paper that you leave on your desk, taped to your monitor, or underneath your keyboard. Secure your passwords as you would your data.

12 Never give your login/passwords to anyone.

13 Use locks on your files and doors that require secured access.

14 Install surge protection and UPS (unlimited power supply) to your computer.

15 Disconnect your computer from the Internet when you're not using it.

16 Make the security of your data a top priority.

EXERCISES

Hashing

Hashing is based upon some very complex mathematical formulas, which are used to produce a unique number (or hash) based upon the contents of an image file. The number produced is unique. In practice, it is mathematically infeasible for two different image files to produce the same hash value. Change one pixel in the image, and the resulting hash value will be completely different from what it was prior to the change. If the file contains header metadata, and that is hashed along with the image, then any changes to the metadata will cause there to be a new hash number.

The hash, or hash value, usually is represented as a string of 32 to 128 hexadecimal digits (the numbers 0 through 9, and the letters A through F), representing the values of the hash's 128 to 512 individual bits.

Hash Definitions

MD5: Message Digest 5—a mathematical formula for uniquely representing any amount of data in just 128 bits (16 bytes).

SHA-1: Secure Hash Algorithm 1—a mathematical formula for uniquely representing any amount of data in just 160 bits (20 bytes).

SHA-224: Secure Hash Algorithm 224—a mathematical formula for uniquely representing any amount of data in just 224 bits (28 bytes).

SHA-256: Secure Hash Algorithm 256—a mathematical formula for uniquely representing any amount of data in just 256 bits (32 bytes).

SHA-384: Secure Hash Algorithm 384—a mathematical formula for uniquely representing any amount of data in just 384 bits (48 bytes).

SHA-512: Secure Hash Algorithm 512—a mathematical formula for uniquely representing any amount of data in just 512 bits (64 bytes).

Hashing Images

There are many hash programs available and a variety of methods of hashing images, but the fastest and safest method is to hash the images as they are being downloaded from the camera or storage card.

One program designed for this is called Authentegrity, a plug-in that works in conjunction with Breeze Systems' Downloader Pro to document the chain of custody and prove the integrity of evidentiary images recorded with a digital camera.

Authentegrity was designed by Erik Berg—a law enforcement professional with more than 20 years experience—and veteran software engineers Carl Ransdell and Matt Brown, in order to meet the unique needs of law enforcement and other professional users of digital cameras who must introduce their images into legal proceedings.

Authentegrity generates a digitally signed text file that contains a wealth of documentary information:

- The IP address for the computer used to download an image from a camera's memory card
- The username of the person logged into Windows when the image is downloaded
- The date and time the image is downloaded
- Two different secure hash values (SHA 256, SHA 512) for each image file downloaded.

Once the plug-in has been installed, it automatically hashes all images that are downloaded using the Breeze Downloader Pro program.

As mentioned, there are many programs available, such as PTHASHER, but they tend to be used after the fact. The images are downloaded to a folder and then a hash value is run.

PTHASHER

This program can be used to compute the hash values for many different types of files, and can also check the hash values created by Authentegrity in Downloader Pro.

PTHASHER can compute up to six different types of hash values: MD5, SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512.

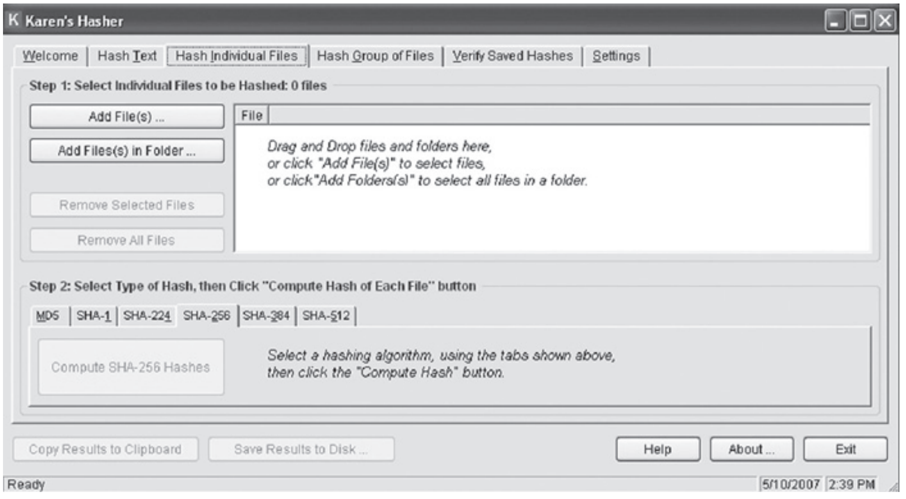


FIGURE 15.E1

Note the tabs at the top of the window in Figure 15.E1. Not only can you hash images, but also text files. It also hashes individual files or a complete folder of images.

The "verify saved hashes" function lets you verify any hash value created by any hash program. It is used to verify the hash values written by Authentegrity in Downloader Pro. Figure 15.E2 provides an example.

FIGURE 15.E2

In Figure 15.E2, note the two hash values created: SHA-256 and SHA-512. To verify the hash values written by Authentegrity, you must hash the file with PTHASHER or another hash program. Refer to Figure 15.E3.

To use PTHASHER:

1. Open the program.
2. Select hash individual files.
3. Add the file or files you want to verify.

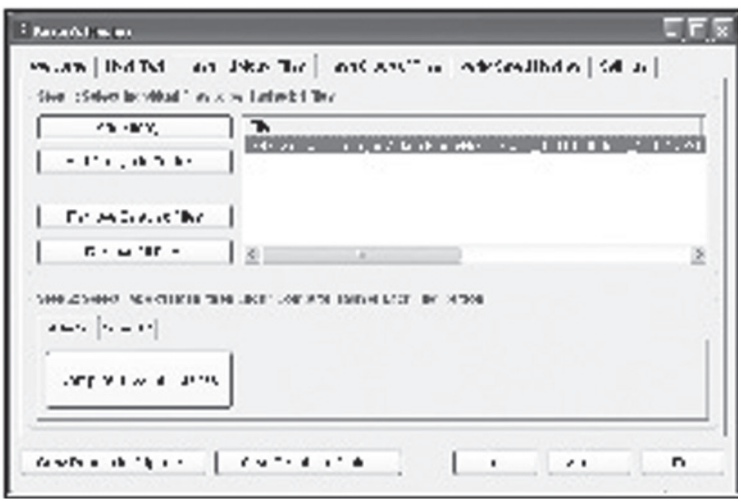


FIGURE 15.E3

4. Highlight the file after it appears in the file window.
5. Select the SHA-256 tab.
6. Select the compute button.
7. Copy the results to the clipboard.

The new hash value is now on the clipboard.

Refer to Figure 15E.4, the text file (hash) created by Downloader Pro. All you have to do is compare it with the original hash value.

Paste the results from the clipboard under the SHA-256 hash value in the text file. Two hash values—one by Authentegrity—one by PTHASHER, are shown in Figure 15E.5.

FIGURE 15.E4

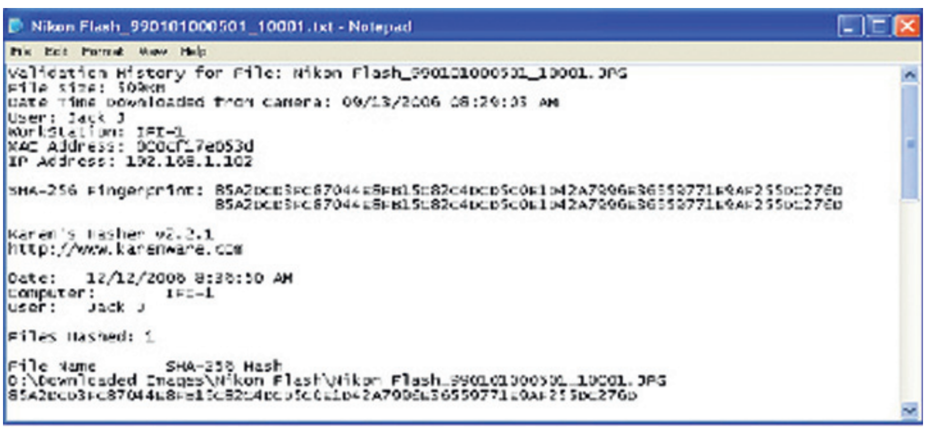


FIGURE 15.E5

Both hash values are the same—one was created at download, the other at a later date to verify the image's integrity.

Another hash program that is fairly popular is File Check MD5. It is a standalone program and does not hash images during the downloading. It is also limited to the MD5 hash values. Refer to Figure 15.E6.

Note: FileCheck MD5 ©2004 Brandon Staggs, all rights reserved. FileCheckMD5 may be freely used and copied but may not be sold. You may place FileCheckMD5.exe on commercial CDs for the purpose of testing the integrity of the CD.

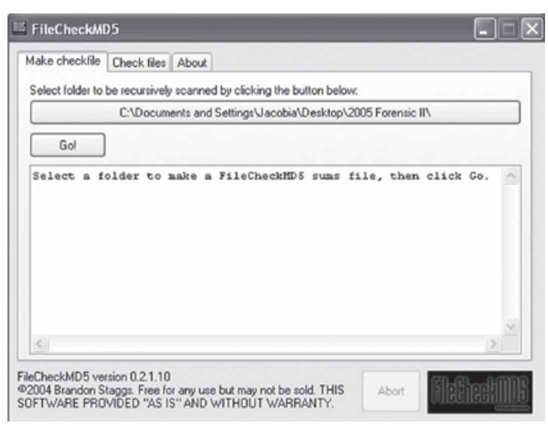


FIGURE 15.E6

This program creates a checkfile (hash). Select the folder where the images are located and MD5 will calculate an MD5 sum for each file. The values (hash) will be saved in a file (FCMD5-sums.MD5) in the same folder as shown in Figure 15E.7.

The MD5 checkfile must remain in the same location as the images. Refer to Figure 15.E8.

Testing is very simple: select the **Check files** tab. Open **Checkfile** and locate the MD5 file you want to verify (see Figure 15E.9).

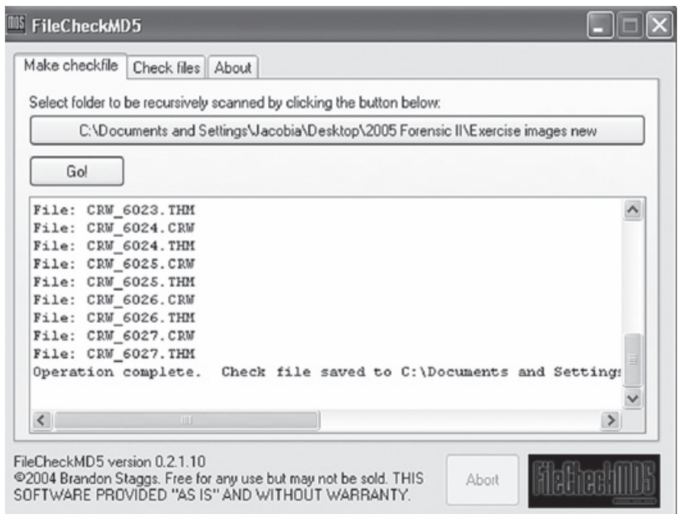


FIGURE 15.E7

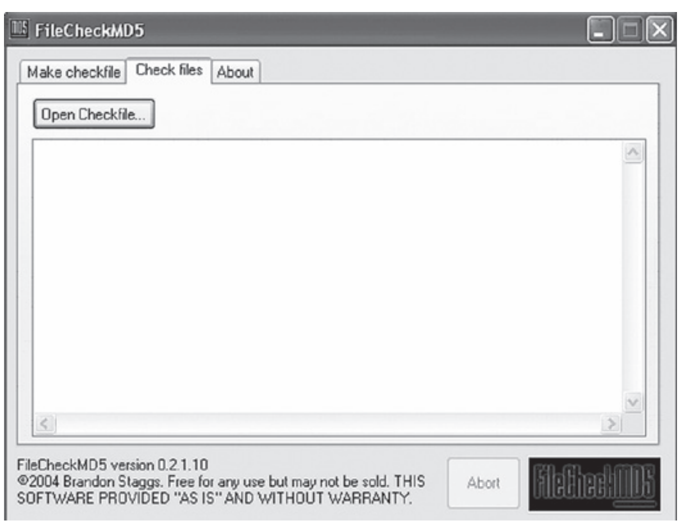


FIGURE 15.E8

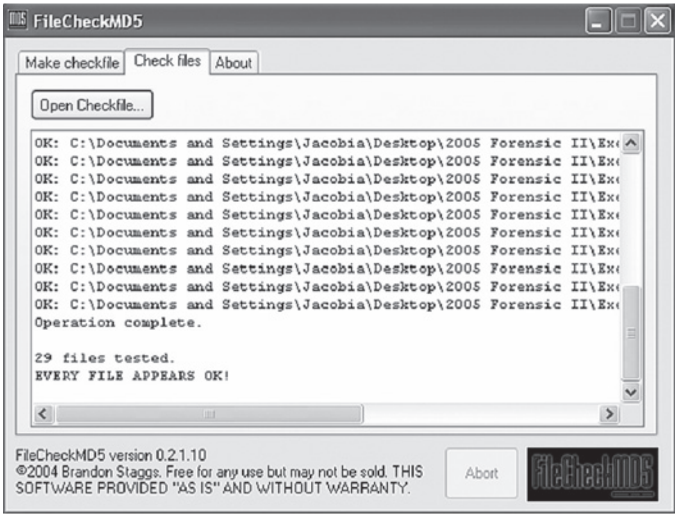


FIGURE 15.E9

The MD5 sums (hash) will be compared. If the files have changed (such as if there was data corruption), the test will show that the MD5 sums do not match.

Verifying Primary Images

Hash values are only one way to verify primary images, but it is not a foolproof method because the images could have been tampered with prior to writing the hash values.

Another check should be the metadata. Although there is software available that will strip the metadata from an image file, it is very hard to alter. Besides, if the metadata had been stripped, it would be a pretty good indication that the chain of custody had been broken.

There are many programs that read metadata, including utility programs that are supplied by the manufacturer. One of the more popular programs is Adobe Bridge, supplied with Adobe PhotoShop.

A lot of information is included in metadata, including all the camera information, creation dates, and modification dates (see Figure 15.E10).

The first item to check is the application section in File Properties. As a rule, if the file is a primary image from the camera, a number will appear usually indicating the firmware of the camera (see Figure 15.E11). The same number will also appear in the Camera Data Software block.

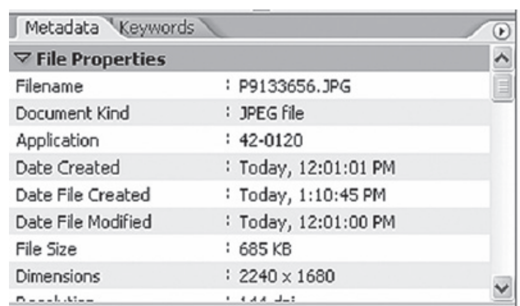


FIGURE 15.E10

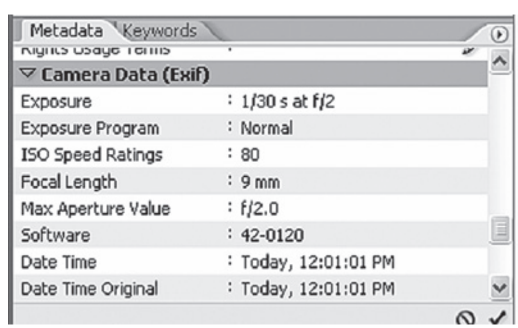


FIGURE 15.E11

This is a pretty good indication that the image has not been opened in a software program and saved. It is not as powerful as a hash value, but is an indication of whether or not someone has worked on the image (see Figure 15.E12).

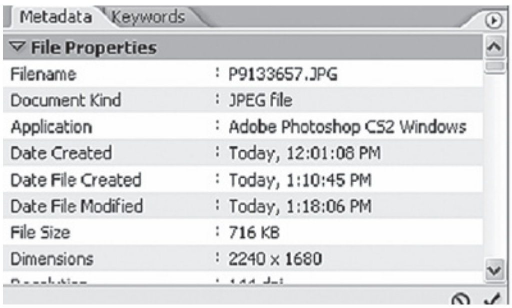


FIGURE 15.E12

If an image has been opened and saved, even using the same filename, the application and software information will change in the metadata. The application information will be replaced by the software used, as shown in the example in Figure 15.E13. Also note the file modified date as compared to that of the original metadata.

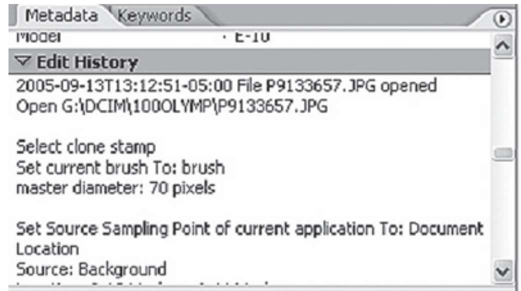


FIGURE 15.E13

Another indication in PhotoShop is the history log. If it has been activated prior to working on the image, it also records in the metadata under the Edit History section as shown in the example.

None of these indicators mean that the image was altered—only enhancements may have been applied. But they do indicate that it is not a primary image "first captured from the camera." In other words, the image has been opened and resaved.

CHAPTER 16

Computing Images

When you take a picture with a traditional film camera, there is no immediate output other than the click of the shutter and the advance of the frame counter. The roll of film has to be completed, removed from the camera, and processed before you can see any pictures. With a modern digital camera, you can see the image on the back of the camera within an instant, giving the impression that the picture goes right from the chip to the viewer and the storage card in the camera. But in reality, the camera had to do a lot of image processing before the image could be viewed; it is just that the processing time is very short.

An *image* is something a person can look at and recognize, and an *image file* is the data needed to produce that image. Image files start out in a very basic form and will have undergone significant processing before they can be used to render a viewable image. The processing has many steps, is very complex, and involves a lot of estimations. The image that is produced is the best estimate of a rendition of the original scene. The processing of film images (as compared to digital) is different in details, but is just as full of estimates and adjustments. Some digital cameras have three sensor chips. Some have the equivalent of the color pixels stacked one above the other. Scanners use separate sensors for the three primary colors, so they can eliminate some of the estimation, but they, too, require significant postprocessing to convert the raw readings into a viewable image. The vast majority of cameras have a single chip with three colors of pixels and a lot of arithmatic is required to produce a satisfactory, viewable image.

The notion of "photographic" accuracy implies that the image is a perfect rendition of the original. This is never the case. The real question is whether it is good enough for us to draw the conclusion enunciated for the image. In other words, is the photo sufficiently accurate to conclude that the footprint that was found was made by the shoe that was found? It behooves the expert giving testimony to have knowledge of the key processing steps and the associated factors that can affect the fidelity of the image.

In the early days of serious digital cameras, the production volumes were very low, and so it was cost prohibitive to put too much processing in the cameras. So the file that came from the camera was not a viewable image—basically, it had not yet been processed. Instead, the camera was sold with software that included an "acquire module," and the raw image file was processed into a viewable image in a computer and not in the camera. As cameras became available for the consumer markets, volumes went up, and it became feasible to include an application-specific integrated circuit (ASIC) in the camera to process the image there. Photographers were thrilled not to have to process each and every image.

Generally these cameras had relatively low resolutions (6 megapixels was the giant for several years), and the cameras could output either a TIFF or a JPEG image—fully processed. But as the number of megapixels went up, the TIFF option became too unwieldy. A 6-megapixel camera will output TIFF images that are over 18 megabytes each, and the number of images that could be put on a single chip decreased dramatically. This gave rise to the RAW file. The RAW file was a throwback to the early days before ASICs in the camera, but photographers did not seem to mind processing each and every image. In fact, many seemed to like it. So today's cameras typically have two basic outputs: JPEG and RAW. Both are significantly smaller files (in megabytes) than a TIFF would be, so more images can be put on a card. However, the capacity of the memory cards is going up so it is likely that the pendulum will swing back and RAW will be supplanted by TIFF once again. In the following sections, the main processing steps will be described in general terms since the specifics are typically proprietary.

SENSOR CHIP DATA

The sensor chip comprises a large number of independent light meters referred to as *pixels*. These count the number of photons that land on them for the time that the shutter is open. They accumulate an electrical charge that is proportional to the total number of incoming photons. This is an analog signal. Each of the pixels has a location given by how high it is in the column and how far across the chip face; and these values, although not yet attached, are numerical (digital). So when the shutter closes, the sensor chip has bundles of electrical charge stored in each pixel location. If the sensor chip is a CMOS chip, it will convert the analog charge value to a number and attach the location data. If the chip is a CCD type, the charge from the various pixels will be pumped out of the sensor chip into another processor chip that will convert the analog values to digital numbers in succession; the order in which those numbers flow from this chip gives the pixels' locations. In either case, at this point, there exists a file containing the basic information needed to make a crude image—but it is not an image yet.

The camera has a number of settings that are set either automatically or by the photographer. In either case, the actual adjustment signals come from a processor chip in the camera that drives all the functions, such as focus, f-stop, shutter speed, and so on. This chip also keeps track of the time, day/date, and other such information pertaining to the image. At its most basic level, the camera reads the chip in the dark to find bad pixels. It reads the overall light level to set exposure, and it reads the amount of red light compared to blue light to determine color temperature (and other functions depending on the make and model). This data is called *metadata*, and is associated with the image information to make up the full image file. Some of the metadata is needed in order to process the image file into an image.

DEMOSAIC

One of the most fundamental steps in computing an image is *demosaic* processing. In most cameras, each of the pixels sees only one of the three primary colors: red, green, or blue. The light-level reading coming from that pixel is only for that one color. So the assemblage of values from all the pixels is in a mosaic pattern, which dictated by the design of the sensor chip. Any one pixel can only indicate the amount of light that fell on that spot that was of the particular color assigned to that pixel. Yet to view an image, three values for each pixel are required—one for each primary. The demosaic process is one that estimates the other two color values for each pixel based on the values achieved in close neighbors.

Prior to demosaic processing, there are three monochrome images, each representing a different color view of the original scene and mixed together in a visually confusing juxtaposition. Also each record has significant spatial gaps since there are pixels that read other colors interspersed in the mosaic. This is not a viewable image and attempting to look at it would be disappointing, to say the least.

Each of the monochrome images has missing values. The demosaic process is similar to interpolating up each of the three separate monochrome images to fill in its blanks. One main difference is that with interpolation, there is no original information about the missing pixel values whereas with the demosaic there is a small amount—that contained in the other color's response (see Figure 16.1). For example, if pixels 5 and 7 in a given row are both red pixels, then pixel 6 is missing red information. Let's assume that the value for pixel 5 is 128 and the value for pixel 7 is 132. We could imagine averaging the values for pixels 5 and 7 to get 130, and then assign this value to pixel 6. Continuing this process across the rows and for all the columns will result in achieving red values for every pixel. A similar process could be applied to the green and blue records to fill in those blanks. The result is that every row now has red, green, and blue values for every pixel. This is an example of a demosaic process based on bilinear estimation. It works in that it yields color images, but it does not give very high quality. Over the years, manufacturers have developed complex mathematics to give very high-quality demosaic processing. These routines are generally proprietary and have a lot to do with the quality of the images that a given camera can produce. But, no matter how complex the math, the basic idea is that the values for two-thirds of the image are estimated.

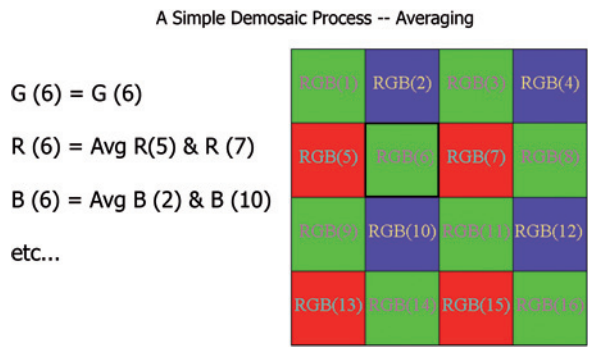


FIGURE 16.1 *Demosaic example. Although most demosaic routines are extremely complex, an extremely simple routine is shown to convey the idea of what such routines accomplish. In this case a simple linear averaging process is shown to indicate that the values of near neighbors are used to estimate missing values.*

NOISE REDUCTION

Each pixel is a light meter. It generally gives a measurement of how much light was absorbed at its location. But, like any light meter, it is not perfect. It could give a reading (albeit a small one) even if no light landed. There could be an imperfect transfer of its charge and conversion to a number. Or it could be oversensitive (or undersensitive) compared to the others on the chip. The first two effects are generally random noise effects that vary from picture to picture. The third tends to persist across several pictures—that is, variable and systematic noise. Figure 16.2 shows the noise in a chip.

The systematic noise can be dealt with by periodically reading the responses of the pixels, making pixel-by-pixel adjustment. For example, if the camera takes a picture in the dark (its shutter closed), all pixels should have very low readings, and any that do not are overly sensitive.

The variable noise must be dealt with as part of the chip design. The transfer of the charge from each pixel must be highly consistent, and the chemistry and level of purity in the chip must be extremely well controlled. It is also possible to apply a smoothing function to the data, but this runs the risk of eliminating some dark shadow signals from the images.

In general, modern cameras have much lower noise levels than their earlier counterparts. Each camera will have some routines specifically incorporated to minimize noise, and these are part of the image processing.

Another noise component will arise from dirt particles accumulating on the surface of the sensor chip. The individual pixels tend to be smaller than 10 microns on a side, and so even a very tiny dust particle can have a significant footprint. Camera manufacturers provide instructions on how to clean the sensor chip, but the overriding requirement is to do it with great care! It is better to keep the camera's insides clean than to have to reach in and clean out dirt.

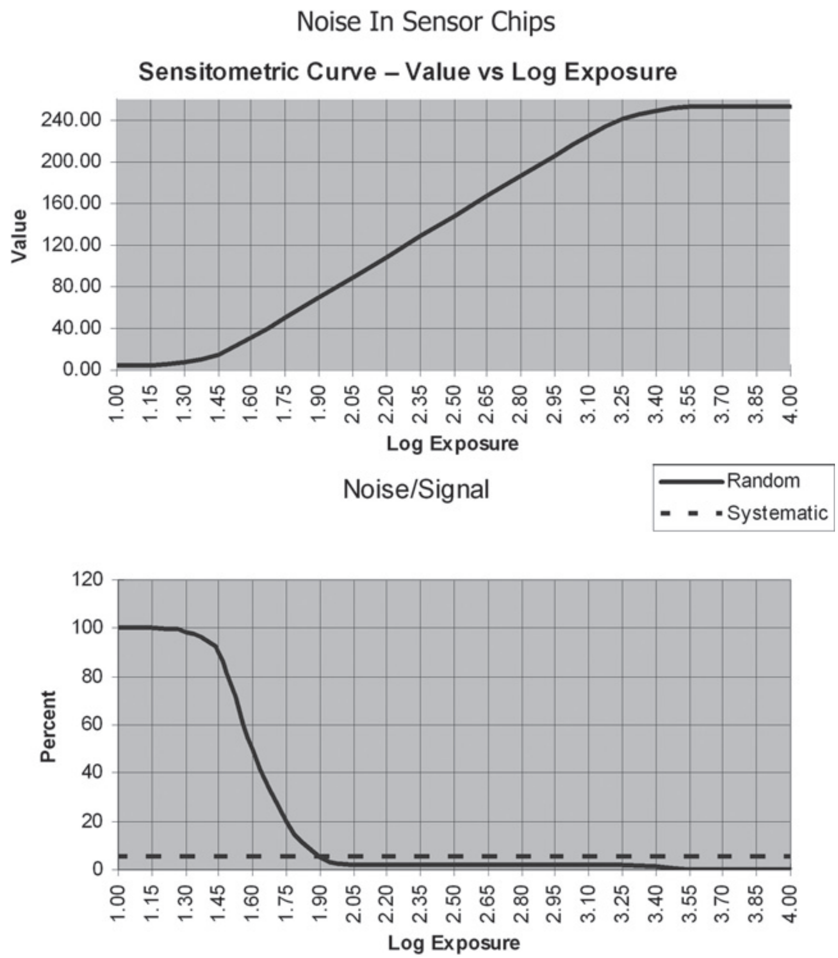


FIGURE 16.2 *Noise in Sensor Chips. The two graphs are hypothetical and show that the random noise generated by pixels is prevalent at low exposure points where the light response is very low. As the light-based signal increases, the relative importance of noise drops off. Systemic noise—for example, dead or overly sensitive pixels—add noise no matter the light level, and this noise pattern is consistent image to image.*

LOGARITHMIC TONE SCALE

The human visual system responds to equal percentage increases in brightness as equal perceptual steps. This is intrinsically a logarithmic scale—equal increments in the logarithm of brightness are perceived as equal visual increments of brightness relative to each other. The pixel light meters on the sensor chip are linear devices. The increase in charge accumulated is equal to the increase in photons absorbed. In order to produce visually satisfying images from the linear sensors, the values must be converted to give a logarithmic scale, as shown in Figure 16.3. This conversion is part of the image processing that is required.

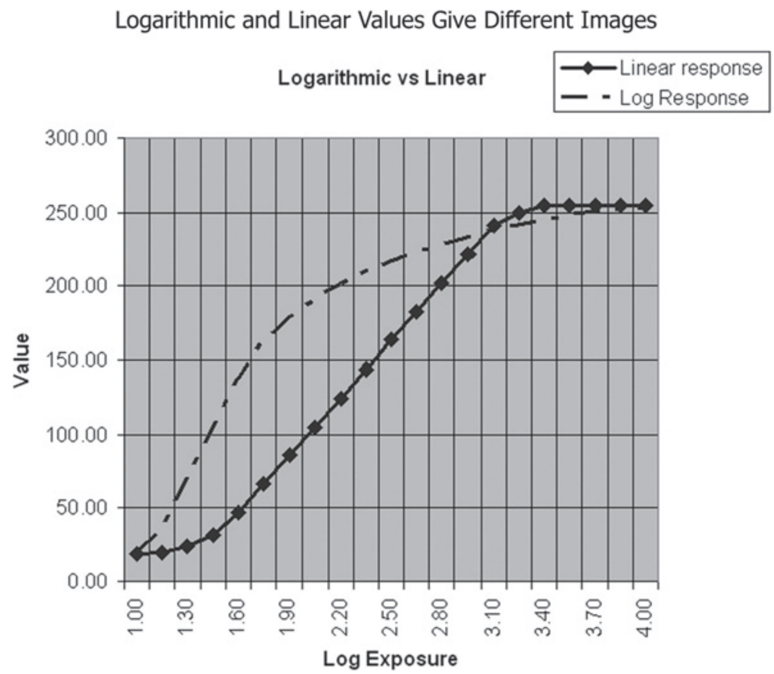


FIGURE 16.3 *Log response vs linear response. The two graphs indicate how different the two responses are. The logarithmic curve is a better indicator of how a person sees images and the linear graph is a better indicator of how the basic light sensors respond to light. The software that computes the image from the raw chip data must make the adjustment.*

ENHANCE IMAGE STRUCTURE

As indicated earlier, the raw data includes gaps. There are spaces between adjacent green pixels, adjacent red pixels, and adjacent blue pixels. Technically, this is called *undersampling* and causes potential errors in reconstructing the image. The demosaic process estimates the values for the missing data. Most of the time the texture in the image is large compared to the gaps, and so the estimates are quite good. Occasionally, though, the detail is of the same order of magnitude as the gaps. In this situation, the estimates might be off by quite a bit. The result is to add color where it should not be. This is called color aliasing and is shown in Figure 16.4. One way to deal with this is to place a ground glass element in front of the sensor chip. This should be designed to spread a small amount of light—about a pixel or so. Then, to counteract the blurring incurred, we must add sharpening to the image. The spreading tends to make the features of the image larger compared to the spaces in the pixel array, and the sharpening filter compensates for that spreading. With newer cameras, the number of pixels has increased dramatically and so the scale of visibility of the problem has decreased. But the basic problem is still there, and new estimates must be made to correct for potentially bad estimates in the first instance.

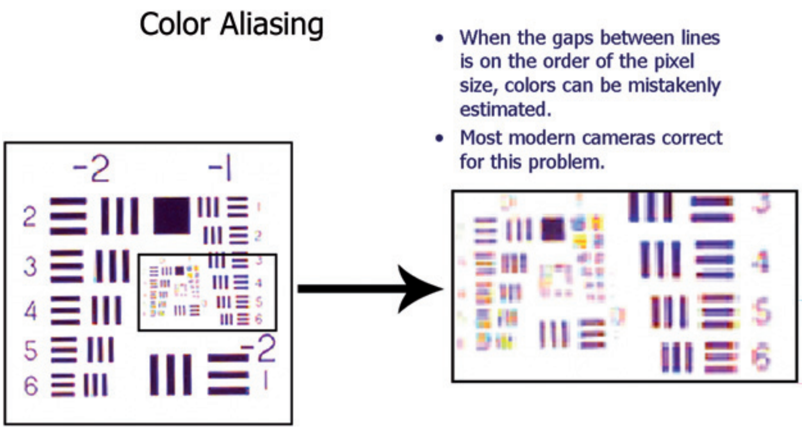


FIGURE 16.4 *Color Aliasing. Under special conditions, usually associated with sharp line edges with high contrast, the demosaic system will incorrectly assign a color value. Typically this is seen as yellow or cyan in areas that should be nearly white. The image includes a rendition of a full US Air Force test chart and an enlarged view of the higher frequency portions.*

Image sharpening can also be used to compensate for the lesser ability of the lens to faithfully transmit very fine detail. In any event, many postprocessing routines include some degree of sharpening.

WHITE BALANCE

If you take a picture of a white card in front of a blue light, the card will be rendered as blue, not white. Similarly the same card in front of a red light will appear red. The card is the same color, but in the two photos it was misrepresented. If the colors were not too saturated, a human observer would note that everything in the blue image seems too blue and everything in the red image seems too red, and that person would make an adjustment for the color of the light. The card would then be seen as white. Daylight is a bit blue as compared to indoor lighting, which is often yellow. But a handkerchief is white in both cases. Cameras are not quite as smart as human viewers and adjustments must be made.

Normal light sources vary in their relative amounts of red and blue light, as just described. Also, normally the brightest thing in the image will be the best indicator of this red-to-blue ratio. Accordingly, if averaging across all the pixels, the brightest ones will provide the bulk of the photons; the average will be dominated by these pixels. Now, if the average for the red pixels is compared to that of the blue ones, the red-to-blue ratio can be estimated. This is a good indicator of the color temperature of the source and can be used to estimate the spectrum of that source.

The general formula for the response of a sensor is given by:

$$R = \int_\lambda (a \cdot \text{Light}) \cdot (b \cdot \underline{\text{Image}}) \cdot (c \cdot \text{Filter}) \cdot (e \cdot \text{Sensor})\, d\lambda,$$

where *Light* is the spectrum of the light source, *Image* is the spectrum of the image component being examined, *Filter* is the spectral nature of any filters in the system, and *Sensor* is the spectral response of the sensor device itself. The coefficients a, b, c, and e are constants and depend upon design parameters and the units of measure. The value for R is what a given pixel returns. The values for the image, filter, and sensor are known before the fact. In order to estimate the value for the image, an estimate of the light source is required. It is here that the color temperature information comes in.

In practice, the readings for red, green and blue will be adjusted relative to each other based on an estimate of the light source. This corrects for the tint that might be added by changes in the color quality of the light source. You can purchase a diffuser that goes over the front of a camera lens and spreads the light from the full scene across most of the pixels. This gives a bit more robust reading of the color temperature since it does not depend on potentially only a handful of pixels. However these tend to be expensive, and most modern cameras are really quite good at this to begin with.

The process just described is referred to as adjusting for white balance. It can be done automatically by the camera, or it can be altered manually in postprocessing. In either case, it is part of the set of estimates that are made in order to get a viewable image from the raw data. In traditional film photography, this is accomplished by the use of filters either on the camera or on the enlarger when making the prints.

Once the proper coloration is determined, the values have to be adjusted depending upon which working space is to be used: sRGB, Adobe RGB, Pro Photo, or another. This ensures that the system will give the best color rendition all the way through the rest of the image-processing and display steps.

FORMATTING THE FILE

With all the adjustments made, the image is now almost viewable if it is a JPEG image. If it is a RAW file, all the estimation processes just described are skipped and will need to be applied in a computer before the image can be viewed. In either case, the information has to be put into a format that a computer or display device will recognize and be able to read. The format indicates where the various elements of the image are located within the file. Standardization is crucial if the image is to be generally available. The file's header will include an indication of the organization and properties of the image. It must indicate how to find and interpret the various pixel values, and it must indicate the bit depth for each value. If compression has been used, the properties of the compression algorithm must be indicated. Since most image files from cameras include metadata, indications must be included regarding where the various metadata elements are to be found. Once these are known, it becomes a readable and viewable image.

If the image is a JPEG image, the parameters of the compression must be included. If the image is imbedded in a RAW file, the metadata is crucial since it will be needed to convert the file into an image.

Bit depth is one issue of concern to many photographers. Modern cameras work with a 16-bit byte. All the data processing will result in round-off errors, and these could accumulate to the point of making a difference in the appearance of an image. Most observers believe that people cannot see any more than 16 million different colors. This can be accomplished with 8-bit bytes. Or, working back from the viewer, the printer or display device needs only 8 bits per color since the viewers cannot see anything beyond this. But the processing can introduce errors, and if there are not enough bits in the original file, these errors can become visible. So the cameras are designed to work with 16 bits. If RAW files are exported, the 16 bits are needed since the processing is done in a computer and not in the camera. In the case of the JPEG images, 8 bits are sufficient since all of the image conversion processing has already been done.

EXPORT THE IMAGE FILE

Most cameras have two ways by which they can export an image. The most common is to put the image files on a memory card that can be taken out of the camera and put into a computer or some other card-reading device. Cameras also have a connector so that they can be connected directly to a computer or printer. Once the method is chosen, the user can take the images from the camera and use them for whatever purpose is intended. Care should be taken to preserve the integrity of the file that was just exported. Once the file on the card or in the camera is taken from that memory, the exported version is all that exists and it is the best record of the image that the camera created. Any further processing of this image should be done on a duplicate of the original and not on the original itself.

SUMMARY

The camera feeds the raw data from the chip (including metadata) to the external device. These data must be converted to a stream of pixel-by-pixel brightness values that represent the layout in the sensor chip. Corrections are made for random and systematic noise.

The separate red, green, and blue sets of pixel readings are then subjected to the demosaic process to obtain all three values for all pixels. An estimation of the color temperature of the light source is made in order to determine corrections for red, green, and blue readings that better represent the objects photographed instead of the light source involved.

The tone scale must be changed from the linear values that come from the chip to logarithmic values that are consistent with human vision. The image structure needs to be enhanced to reduce color aliasing and add sharpening.

The image information has to be packaged in a particular file format in order to ensure that the image can be opened and viewed in an external device. If the image is rendered as a JPEG, then it will probably have 8 bits per byte. If it is rendered as an unprocessed, RAW file then it will probably be in 16 bits per byte so that the external processing needed to create a viewable image does not incur excessive round-off errors. Finally the camera exports the image file either on a removable storage medium or via a direct wire connection to a computer or printer. The file should include all the metadata and it *must* do this if it is a RAW file.

A RAW image file is sometimes considered a digital negative since it cannot reasonably be created outside of a camera. There simply is no software to allow this.

S

EXERCISES

Raw Images

There is probably more misinformation, disinformation, and lack of information about RAW files than any other digital imaging topic.

Recording the Image

Depending on the camera's circuitry, either 12 or 14 bits of data are recorded. If the camera records 12 bits of data, then each pixel can handle 4096 brightness levels. If it records 14 bits, it can record 16,384 different brightness levels. The information can then be exported in 16-bit mode. The 12 or 14 bits recorded by the camera are then spread over the full 16-bit workspace.

A raw image file is a record of the data captured by the sensor. There are many different ways of encoding this raw sensor data into a raw image file, but in each case the camera records the *unprocessed* sensor data. Information about the camera's settings at the time the picture was taken and the camera's estimate of the color temperature are also recorded.

Most digital cameras have two choices for image format: JPEG or RAW. The difference lies in the compression of the image and the way the image is processed by the raw converter.

Putting it simply:

- A JPEG image is processed by the raw converter in the camera using all the settings from the camera. Then it is then compressed.
- A RAW image is compressed in the camera using lossless compression. Later it will be processed in a software program supplied by the manufacturer using the settings applied in the software. The user can use different settings, however.

JPEG versus RAW

JPEGs offer fairly limited editing headroom. If a significant degree of compression is applied, large moves in tone and color tend to exaggerate the 8×8-pixel blocks that form the foundation of JPEG compression. And although JPEG does a decent job of preserving luminance data, it applies heavy compression to the color data, which can lead to issues with skin tones and gentle gradations when trying to edit the JPEG.

In RAW format, however, we get unparalleled control over the interpretation of the image. The only on-camera settings that have an effect on the captured pixel values are the ISO speed, the shutter speed, and the aperture setting. Everything else is under the user's control when converting the RAW file into an image. We can reinterpret the white balance, the colorimetric rendering, the tonal response, and the detail rendition (sharpening and noise reduction) with a great deal of freedom.

There are two basic reasons for using the RAW format:

1. It doesn't compress the primary image:
   - **a** We can export directly from RAW to TIFF in 16-bit mode.
   - **b** This may involve lossless compression.
2. It allows for fine control over the adjustments to the image:
   - **a** We can make all the adjustments using the raw utility.
   - **b** The adjustments are made to a 16-bit image, allowing more accurate control over the adjustments to the colors and densities.

The disadvantages of RAW format are:

- It is very time consuming, because each image must be opened and enhanced individually.

- It works well for individuals adjusting their own photographs, but if the user didn't take the photograph him- or herself, would he or she know how to adjust the colors or the densities?
- The photographer should not archive a RAW format since the conversion utilities are proprietary and unique to the camera manufacturer or software vendor. They may not be supported for the decades of storage that might be required. Eventually RAW files should be converted to TIFF or JPEG.

The advantages of the RAW format are:

- The image is recorded in 16 bits, creating many more colors—approximately 281,474,976,710,656; or 65,536 gray levels. JPEG records only 8 bits, approximately 16,777,216 colors.
- The image adjustments to the color balance, general levels, and such are made using the raw utility. This does round off in much smaller increments, thereby making them more accurate than adjustments made in 8 bits.

Conclusion

If you choose to use RAW file output, make as many adjustments as possible in the raw utilities before exporting the image to the image-editing software. Do not have another person process the images. If the adjustments are not done in the raw utility, the purpose of using the RAW format will have been defeated.

Working with RAW Files

The images for this exercise are proprietary Canon Raw files. Like any raw file, there are basically two ways of opening them. You must either have the viewer utility program provided with the cameras, or Adobe Photoshop with the correct plug-in. The example in the exercises demonstrate the tools provided in the Canon utility program, but the same adjustments, and many more, are available in PhotoShop.

The images provided on the website are:

CRW_5852
CRW_6023
CRW_6024
CRW_6025
CRW_6026
CRW_6027

Other associate information for each image is in the folder. Download it all.

Use these images to explore the enhancements available in the Canon viewer utility of Adobe PhotoShop.

File Viewer Utilities

RAW images are processed so that they are optimized for their intended use. In addition to using the same white balance and sharpness settings as those set on the camera, File Viewer Utility lets you specify additional processing parameters, such as digital exposure compensation and white balance adjustment using the Click White Balance function.

As an example, we will use the Canon File Viewer Utility for the EOS 10D camera. The File Viewer Utility allows users to view RAW images taken with the indicated Canon camera, enhance the images, save those images as files, and transfer the files to applications (see Figure 16.E1).

FIGURE 16.E1

Most file viewers will display a histogram of the selected image and the image information (metadata).

- The histogram displays the values for all the pixels in the image.

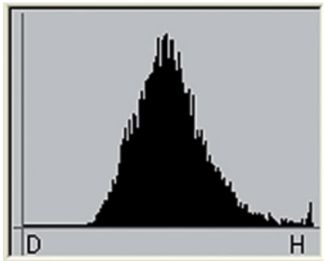


FIGURE 16.E2

- The metadata displays information that includes the filename, camera model, date and shooting conditions, among other data.
- Image information can be copied and pasted to a word-processing program, and then printed.

Toolbar

Each utility program will have a variety of tools available for editing the selected image. An example of the Canon tool bar is shown in Figure 16.E3.

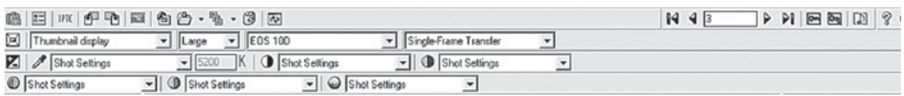


FIGURE 16.E3

Preferences are available in each program, as shown in Figure 16.E4.

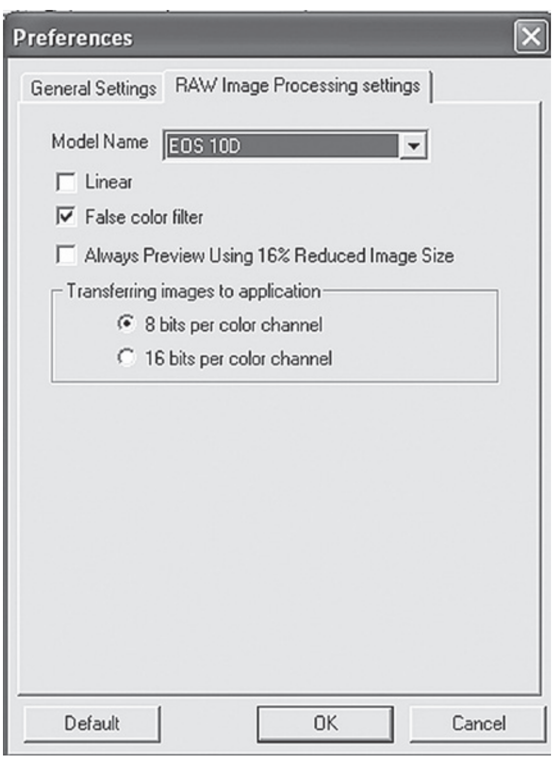


FIGURE 16.E4

- **Image transfer application**. This lets you specify the application software to which images are transferred when the Transfer Images menu is selected. Specify an executable application software file that can open JPEG, TIFF, and other image file formats.
- **Image transfer folder**. When transferring an image to an application, the image file is generated in the Image Transfer folder specified in this field, and then this image file is opened by the specified application.

Most programs will require you to set up how the RAW image will be processed. Some of the settings may include:

- **Linear**. When this checkbox is selected, linearly converted images are transferred and saved when transferring 16-bit images to an application or when saving TIFF (16-bits/channel) files.
- **Always Preview Using 16% Reduced Image Size**. When this checkbox is selected, RAW image previews are displayed at a 16% reduced image size.
- **Transferring Images to Application**. This sets the bit depth (choice to transfer color at either 8 bits or 16 bits) for converting and transferring RAW images. If transferring 16-bit images, the application receiving the images must be able to support 16-bit color.
- **Also Get JPEG Image**. Some cameras have a setting that allows the outputting of both a RAW and a JPEG file. When this has been done and this checkbox is selected, both a JPEG and a RAW file transferred to the application at the same time.

Working on the RAW Images

RAW images taken on most digital cameras (that support RAW capture mode) can be converted to common file formats, such as TIFF or JPEG, using development parameters set in File Viewer Utility.

Follow these steps to adjust the quality of a RAW image, to convert it, and to save it.

Common adjustments available in a raw utility are:

- Digital exposure compensation
- Adjusting the white balance—this can usually be done by sampling a white point with the eyedropper or by selecting a light source (see Figure 16.E5).

FIGURE 16.E5

- Contrast adjustment
- Color saturation adjustment
- Color tone adjustment
- Sharpness adjustment
- Specify the color working space as in Figure 16.E6.

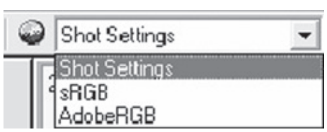


FIGURE 16.E6

When the adjustments are finished, the file will normally be converted to a different format for archiving. The file formats available for conversion will depend on the program.

The Canon utility has three conversion options, as shown in Figure 16.E7:

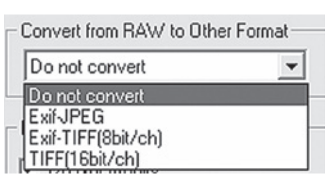


FIGURE 16.E7

- JPEG
- TIFF (8-bit)
- TIFF (16-bit)

Four compression levels are available when converting to JPEG images. Always convert forensic images using minimal compression.

Note: Saving as (CRW) means "Do Not Convert", and does not save any of your adjustments—it saves only the raw RAW file. File sizes vary for each option, and are shown in chart in Figure 16.E8. Compare the file sizes in the chart.

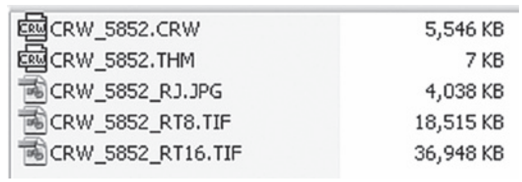


FIGURE 16.E8

Extracting and Saving JPEGs

This feature is only available for RAW images taken with some cameras.

An extracted JPEG image is a JPEG image that was embedded in a RAW image. Extracting JPEGs from the RAW images makes it possible to easily generate JPEG images that can be viewed in other software without having to develop the RAW images. There are many programs that will open raw files from different cameras, but the controls for enhancing the images vary tremendously. Most camera manufacturers recommend that you use their utility program when working with raw files.

Adobe PhotoShop has plug-ins for most of the high-end cameras and will allow enhancement of the RAW images. Note that the plug-ins were developed by Adobe and are not necessarily endorsed by the camera manufacturers.

Camera Raw Plug-in

Although each camera uses a unique format to save the camera raw image data, the Photoshop Camera Raw plug-in can open many cameras' raw file formats.

Opening a camera raw file in Photoshop generally opens the Camera Raw dialog box. It's possible to suppress the dialog box from displaying for each file when opening a batch of camera raw files.

The histogram shows all three channels (red, green, and blue) of the image simultaneously. The histogram updates automatically as you adjust the settings in the Camera Raw dialog box. The image adjustments available are:

Color space. Specifies the *target* color space profile. Generally, this should be set to the same value as your Photoshop RGB working space (see Figure 16.E9).

FIGURE 16.E9

Pixel bit depth. Specifies whether the image opens as 8 or 16 bits per channel in Photoshop, as shown in Figure 16.E10.

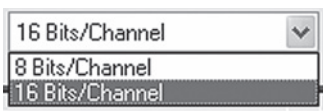


FIGURE 16.E10

Size. Specifies the pixel size at which to open the image. The default for this setting is the pixel size you used to photograph the image (see Figure 16.E11).

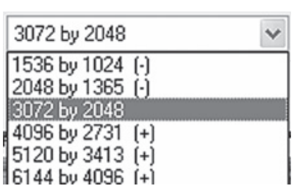


FIGURE 16.E11

Resolution. Specifies the printing resolution or the amount of data in the image once the number of pixels and the bit depth have been chosen (see Figure 16.E12).

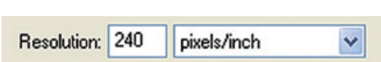


FIGURE 16.E12

White balance control. A digital camera records the white balance at the time of exposure as a metadata entry. This is read by the Photoshop Camera Raw plug-in and set as the initial setting when opening an image in the Camera Raw dialog box (see Figure 16.E13).

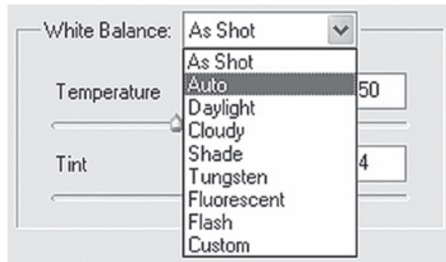


FIGURE 16.E13

Correcting the white balance. Moving the Temperature slider adjusts the color temperature being selected for the light that was used. The tint lets you fine-tune the white balance to compensate for green or magenta tint in photos (see Figure 16.E14).

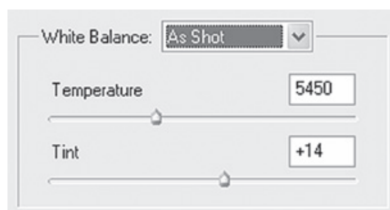


FIGURE 16.E14

Exposure. Adjusts the brightness or darkness of the image. The values, shown in Figure 16.E15 are in increments equivalent to f-stops.

FIGURE 16.E15

Shadows. Controls what input levels will be mapped to black in the final image (see Figure 16.E16). Using the Shadows slider is similar to using the black point slider for the input levels in the Photoshop Levels command.

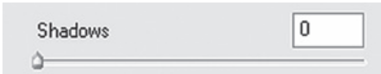


FIGURE 16.E16

Brightness. Adjusts the brightness or darkness of the image, similar to the Exposure slider. However, instead of clipping the image in the highlights or shadows, Brightness compresses the shadows and expands the highlights when the slider is moved to the left (see Figure 16.E17).

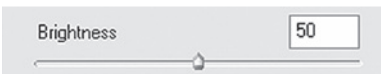


FIGURE 16.E17

Contrast. Adjusts the mid-tones in an image (see Figure 16.E18). Higher values increase the mid-scale contrast, whereas lower values produce an image with less contrast.

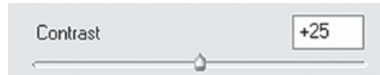


FIGURE 16.E18

Saturation. The control depicted in Figure 16.E19 adjusts the color saturation of the image from –100 (pure monochrome) to +100 (double the saturation).

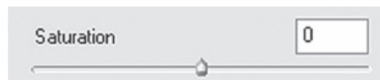


FIGURE 16.E19

CHAPTER 17

Establishing Quality Requirements

Artistic photographs are intended to provoke an emotional response. Forensic photographs are intended to provide a basis upon which an expert witness can develop an opinion regarding evidence, or help the finder of fact (the judge or the jury) see what the original scene was like. In both cases, it is the content of the image that matters, not the technical details of how the photography was accomplished. So the important requirements come down to the ability of the photo interpreter to render a valid and reliable opinion. If there are image features that are not relevant to rendering the opinion, they do not matter one way or the other. If there were features of the original object that are not reproduced well enough to allow the viewer to draw valid and reliable opinions (as an expert witness), then the image is inadequate. The ultimate image-quality requirement is that the system be able to reproduce the features needed for experts in the field to render valid and reliable opinions. If the smallest feature that is needed is the size of a grain of sand (about 0.5 mm), and if the observer needs 10 brightness levels' difference to see a feature against a background, then these are the requirements. Consider two systems: the first is able to render 0.1 mm features and can yield 15 brightness levels of difference; the second is able to render 0.03 mm features with 12 brightness levels of difference. Both of these would be equally as good for the given application. Making the technical measurements is not nearly as difficult as determining what the examiners need. This can be done only by testing their ability to perceive key features.

PERFORMANCE MEASUREMENTS

There are many dimensions in which one can compare photographic systems. These might include the time from exposure to viewing, color gamut, cost, permanence of stored images, optical distortion, color fidelity, tone scale accuracy, and so on. Although many of these are important, they do not vary much among the realistic contenders for forensic imaging applications. When it comes to most forensic applications, though, the key dimensions come down to resolution and dynamic range. These translate into the questions of: Can I see the finest detail that I need to see? and Can I record both the light and dark areas of the types of scenes that I encounter? As a result, any experiment designed to measure performance requirements must clearly address these two dimensions. Others are nice to have, but these are essential.

AN EXAMPLE OF DETERMINING REQUIREMENTS

In 2005, the Institute for Forensic Imaging (Effect of photographic technology on quality of examination of footwear impressions; H. Blitzer, R. Hammer, J. Jacobia; *Journal of Forensic Identification* **57**(5), September/October, 2007) conducted a study to determine one set of application-specific requirements. This experiment addressed how well-experienced footwear examiners could see features in photographs of high-resolution footwear impression photos. Each examiner received a set of eight hypothetical cases. Each case had an exemplar image and crime scene photo. The crime scene photos were made using a variety of photographic systems. The exemplar had five marked areas. Each examiner was to rate how well he could see any features in marked areas on the scene photo. The visibility ratings were made using a scale that ranged from 1—clearly not there to—5, clearly there. For each examiner and for each case, we were able to generate a score that indicated how well the examiner felt the feature was or was not reproduced. The case images were created in a variety of ways. There were two digital cameras (one with 6 megapixels and the other with 14 megapixels), two digital printers, 35 mm film, and 120-sized film. Film negatives were printed on traditional photographic paper. Also, four different shoes were used. Forty-eight anonymous examiners completed the project.

When all the data was studied, the results showed that 120-sized film gave better results than all the alternatives. This is a statistically significant finding. All the other systems were not statistically different from each other. One of the digital printers gave generally better performance than the other, but the statistical significance of the finding was just short of the threshold.

What the study shows is that examiners reviewing images in their normal mode of full-sized print comparisons did not see the differences among most of the options when making footwear evaluations. This finding may not extrapolate to fingerprint comparisons; a separate study would be needed to determine that. It also probably does not apply to on-screen comparisons. Although not a part of the study proper, casual viewing of images on a computer screen make it clear which images were made with the 14-megapixel camera and which were made with the 6-megapixel camera. Again, a separate study would be needed to confirm the difference as meaningful to examiners, and only then would the higher resolution become a requirement. The findings are summarized as follows:

Camera	Printer	Score	Different?*
120 Film	Chemical Paper	4.57	Yes – from all others
14 mgpixels	Noritsu	4.24	No
6 mgpixels	Noritsu	4.16	No
35 mm film	Chemical Paper	4.09	No
14 mgpixels	Desk Inkjet	3.96	No
6 mgpixels	Desk Inkjet	3.93	No

**Statistically significantly different at the 95% confidence level.*

Small, variable, and uneven sample sizes limited attachment of significance to several potential conclusions, but overall the recommendation is:

- Use medium format film on high-resolution impressions where the full impression must be captured in a single photograph.
- Use a digital camera with 8 or more megapixels for all other situations.
- Take separate photographs of the heel and sole portions when very high resolution is needed and two separate photos is okay.
- Choose a very high-quality printer for digital pictures.

In summary, the protocol for determining requirements is as follows. Testing is required to determine image quality requirements for a specific discipline:

1. Several photos of representative objects are made covering a range of image-capture quality levels.
2. The images are studied by a set of examiners in that discipline and they are asked to make the kind of analyses that they normally make.
3. The ability of the image to support the normal tasks is measured and each image is given a score.
4. The average scores for the images at each quality level are an indication of the degree to which that level of quality is necessary to that set of users.
5. The original set of images should be structured to show the image quality factor (e.g., resolution, dynamic range, etc.) that is the main contributor to the requisite quality.

COMPARING FILM AND DIGITAL SYSTEMS

In the best practices documents issued by the Scientific Working Group on Imaging Technology during years 1998 to 2005, film photography was recommended for most applications and digital photography was suggested as an addition to the film-based record. Those documents are now being updated to recognize the dramatic improvements in digital photography in the past several years. Yet the question remains, is it acceptable to change from film-based photography to digital photography for forensic work? The answer will depend on the specific application, but it is possible to show the relative performance levels in general terms. The key attributes are:

- Resolution—the ability to see separate items in the object as separate in the image
- Dynamic range—the input range of light levels over which the system is responsive

The remainder of this chapter deals with comparison studies conducted at the Institute for Forensic Imaging. The results were presented by Herbert Blitzer at the International Association for Identification Educational Conference in July 2007.

RESOLUTION IN DIGITAL CAMERAS

There are several factors that can affect the resolution that can be obtained from a given camera. The external factors include the selection of the lens and the sharpness of the focus setting. For this exercise, a high-quality lens is assumed so that the impact is minimal. Also it is assumed that the camera is in sharp focus. Internal factors include the design of the specific sensor chip, the sampling and demosaic processes, and color-aliasing correction. Experience shows that these are all much less of a factor than the universal pixel conversion efficiency. This concept was introduced in Chapter 6 but will be revisited here. The pixel conversion efficiency is the number of adjacent pixels in a row required to resolve a vertical line pair. A line pair comprises one dark line and one light line in parallel. The pixel conversion efficiency is derived from the geometry of the formation of an image in a digital camera.

To develop the concept, consider a test target that is two parallel dark lines on a light background. The space between the two dark lines is equal in width to the width of each of the dark lines. One of the dark lines plus its adjoining light line is a line pair. The test target can be held close to the camera, a moderate distance from the camera, or far away. The lens will be refocused in each case so that a sharp image is created.

In the back of the camera, the sensor chip has discrete pixels, which we will assume are in a rectangular array. The image of the test target is focused on the sensor chip. When the test target is close to the camera, the lines are

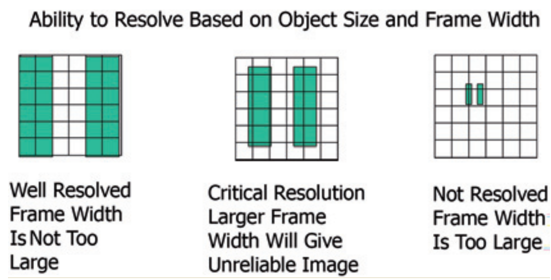


FIGURE 17.1 *Resolution Based on Object Size and Frame Width. The depiction on the left shows an image of a two-bar pattern superimposed on a sensor chip. There is a small frame width and so there are four pixels per line pair. Clearly the object will be well resolved. In the center is the critical frame width, yielding three pixels per line pair. At the right is a situation where the frame is too wide and the bars will not be resolved at all.*

wide compared to the width of the pixels and there will be multiple pixels across each line. The pixels will have no trouble seeing the individual lines. This is shown on the left in Figure 17.1. When it is far away, the lines are small compared to the pixel widths. Now it is clearly impossible for the lines to be distinguished. This is shown on the right in the figure. Somewhere between the close and distant settings, there is a critical setting where the lines will just be resolved. It turns out that we need to have at least three pixels just to resolve a line pair to achieve this condition. You might assume two—one for a dark line and one for a light one—but if the target lines are not carefully lined up with the pixel, this will result in a blur. With three pixels per line pair, it will always be possible to distinguish the two lines. They might not be well formed, but they will be resolved as separate lines.

Consider the Nyquist criterion, which states that at least two samples are required per wave length (two pixels per line pair in this case) to resolve a wave. But this does not apply in photography. The Nyquist derivation requires the use of the sinc function (sinX)/X to reconstruct the sampled wave, and this function is not used in image formation. In other words, Nyquist does not apply! Instead, as has been demonstrated, it takes at least three pixels to reliably resolve a line pair. We refer to the number of pixels required to resolve a line pair, N, as the *pixel resolution efficiency*. As we have seen, it is normally such that N = 3. With low-quality lenses, poor demosaic processing, and some other factors at adverse values, N may be greater than 3, but it will not be less.

To see the N = 3 criterion, examine Figure 17.2, which shows an image from a digital camera where the test target, which had 10 lines, is very near the critical distance. Note that the target was not perfectly aligned with the rows and columns of pixels in the camera. It would be extremely difficult to achieve such an alignment. Nonetheless, it is clear that there are three pixels for each line pair, and the 10 lines, although not well formed, are distinguishable.

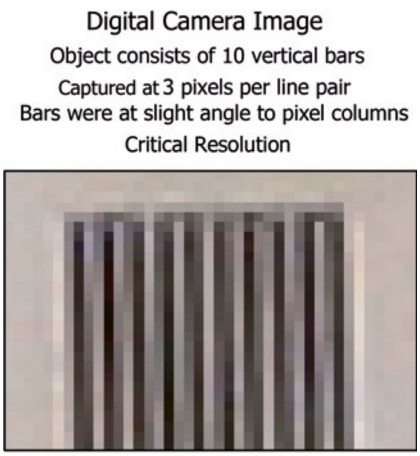


FIGURE 17.2 *Digital Camera Image. Shown is a crop from a photo taken at just about the critical frame width. There were 10 bars on the chart and we can easily count the 10 bars in the image. The diagonal distortion was due to the fact that the bars and the pixel columns did not line up exactly.*

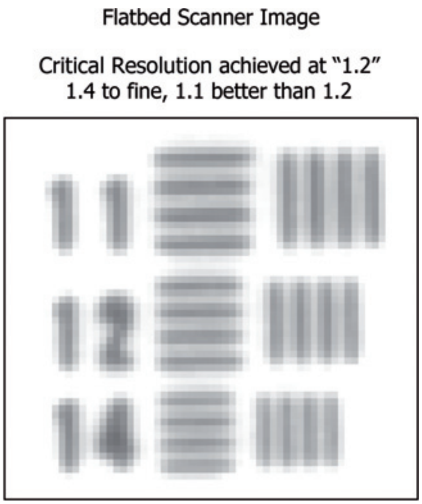


FIGURE 17.3 *Flatbed Scanner Image. Shown is a crop from a flatbed scanner rendition of a four-bar resolution chart. Portion 14 shows the N = 3 condition and the bars are resolvable. The two portions above, 12 and 11, are at greater than N = 3 and the bars are better resolved.*

Figure 17.3 shows an image from a flatbed scanner. The N = 3 condition is shown as 14 at the bottom of the figure, and we can see that the lines are just barely resolvable. The larger targets, shown as 12 and 11 above, are more clearly distinguishable and better formed. The upper two have just over three pixels per line pair.

Although there are several factors that can limit the resolution to some degree, the dominant effect in modern cameras tends to be N, pixel resolution efficiency. Once it is known that it takes three pixels to resolve a line pair, it is easy to determine the level of object detail that will be resolvable as a function of the width of the object frame (or distance from the camera with a given lens). The concept will be developed using the width dimension, but the same argument can be followed using the height dimension.

The sensor chip in the camera has a fixed number of pixels across its width, P_w. If the picture to be taken is of a letter-sized sheet of paper (11 inches across), and if the paper takes up the full width of the frame, then there will be P_w pixels per 11 inches. If the sheet of paper is 22 inches, then there will be P_w pixels per 22 inches, and so on. If the width of the frame at the point of focus is called W_o (width of the object), then in general, there will be P_w / W_o pixels per unit of linear measure (inches in the preceding). Since it takes N = 3 pixels to resolve a line pair, the line pairs per unit of linear measure will be:

$$L_{pm} = P_w/(W_o * N)$$

This same equation was shown in Chapter 6, but it was stated in terms of the size of the smallest details (a pair of lines), or D_o

$$D_o = (W_o * N)/P_w$$

There are a few important aspects of these equations that are worthy of note. First of all, the ability of the camera or scanner to record fine detail is rated *on the object*. In film systems, resolution traditionally refers to the size of detail *on the film*. When considering the detail on the object, *bigger frame widths result in lower resolution*. Conversely, bringing the object closer to the camera and reducing the frame width will result in the ability to see the finer details of the object. This is eminently intuitive. Likewise it is intuitive that if the camera has more pixels, its ability to resolve details at any given frame width will be better. Lastly the focal length of the lens is not in the equation. This seems counterintuitive at first, but the ability to see detail on the object is primarily a function of the frame width and the number of pixels and not the focal length. If the photographer wants a greater or smaller frame width, this can be done by changing lens focal length and/or distance to the object. In any event, the effect of the focal length of the lens will be cancelled out by distance to the target.

The final result is that if you want to see more of an object's detail, move it closer to the camera. To determine how close, use the previous equation and the lens formula. If the reduction in the frame size makes it so that the full object cannot fit into the frame, and if that is required, you must get a camera with more pixels. Those are the only options.

RESOLUTION IN FILM SYSTEMS

Traditionally, film resolution will be rated either in terms of modulation transfer or line pairs per millimeter *on the film*. This makes sense since the film type can be used in a variety of sizes and cameras, and so a measurement of the film per se is very sensible. As was shown, with digital cameras (where the chip is an integral part of the camera), we are interested in the full camera's ability to resolve fine detail in frames of different sizes. Accordingly, the first step in comparing film-to-digital system performance is to extend the film's practical capability to a frame width. The next problem is to use a concept of equivalent pixels, since films do not have actual pixels. With these steps done, it is possible to evaluate a film-camera system in terms of equivalent pixels. But it does not end there. Most of the time, with film, you do not look at the image on the film, but rather at prints made from the film. The notable exceptions are when you look at a slide via a slide projector, or if you scan the film and look at the image on a computer screen.

The place to start is with the film. Film manufacturers rate their films in line pairs per millimeter, and it is easy to look up the data on their web sites. For example, the Eastman Kodak Company shows that Kodak T-Max 100 film has a resolution of 200 line pairs per millimeter. To get this value, they use a very special test target. It is a sheet of optical-quality glass with a very thin coating of metal on one side. A pattern has been etched into the metal using a type of photolithographic technique that is similar to that used to make integrated circuits. The film is put directly against the test target, and the combination is put into a vacuum frame to ensure intimate contact. Finally, the film is exposed through the test target and processed under carefully controlled circumstances. In the studio of the Institute for Forensic Imaging, using good-quality studio circumstances and a top-of-the-line 35 mm film camera, a Canon EOS-1 with a 50 mm Canon EF lens, a maximum value of 42 l/mm was achieved. Figure 17.4 shows one of the test shots, which has both a series of bars and a bar chart. At the bottom of the figure are reproductions of the image from the film on the left and the print from that negative on the right. The losses are immediately evident. Since the size, on the film, of a 35 mm frame is 24 × 36 mm, this would result in 1008 × 1512 line pairs per frame.

So far, the process is fairly straightforward. To go further, it is necessary to bring in equivalent pixels. To do this, we must go back to the pixel conversion efficiency demonstrated by digital cameras and scanners. This argues that it takes three pixels to resolve a single line pair. So, when multiplying by 3, the film in the camera is behaving as if it has 3024 × 4536 pixels, which comes out to a total of 13.7 megapixels. This implies that the 35 mm camera with the selected film will behave as if it were a 13.7 megapixel digital camera. Further tests in the laboratory confirmed this and, in the process, make legitimate the notion of three pixels per line pair. It is interesting that if the analysis is run using the value of 200 l/mm, it would be the equivalent of 311 megapixels. So film has a very high native capability, but in practical terms, a 35 mm camera can be expected to act similarly to a 14-megapixel digital camera.

If the images are to be viewed on a computer screen, no further analysis is needed. Scanners are fully capable of capturing the fine detail on the film.

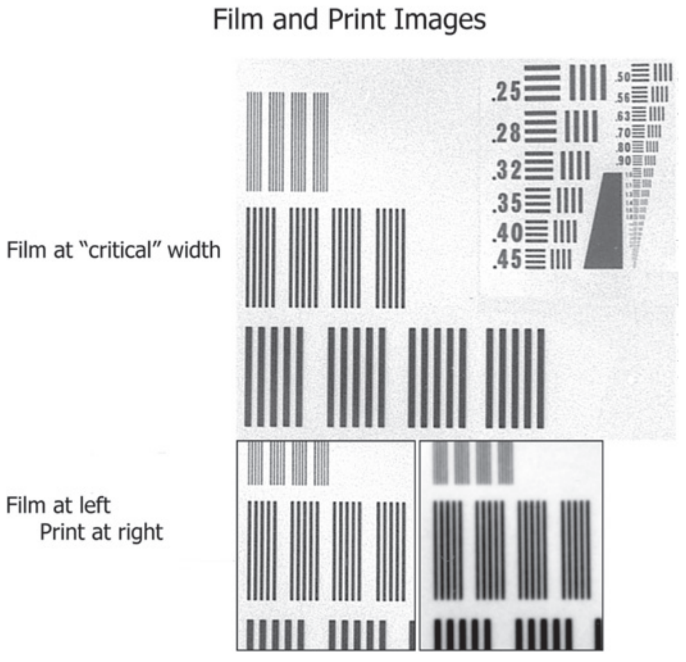


FIGURE 17.4 *Film and Print Images. Shown are a film image at critical frame width; the five-bar charts are clearly defined. Specific resolving power was taken from the four-bar chart in the same image. At the bottom of the figure are shown a crop from the image on the film and the same image from the resulting print. The losses are quite apparent.*

However, if a paper print is to be made, we must consider the losses incurred in this process. Our testing showed that losses can be as high as 25% in each direction. That is, the film image has 3024 × 4536 equivalent pixels, but after losses, this reduces to 2268 × 3402, or 7.72 equivalent megapixels. In other words, if the film camera is to be used in conjunction with a print-making process with 25% loss, the camera-print system will act as if it were similar to an 8-megapixel camera.

The losses come from two factors. One is the native resolution of the print paper and the other is the capability of the enlarger and its lens. Testing showed that we were getting about 7.5 l/mm on the paper itself. This is reasonable compared to the fact that human vision is limited to 5 l/mm at a normal 50 degree viewing angle. If the 7.5 l/mm is extrapolated up to the print size, it is a limiting factor in small images and not a limiting factor in larger ones. A 4 × 6″ print would not be able to hold more than 3.5 megapixels, but a 240 × 360 millimeter (9.4 × 14.2″) print—which might be needed to accommodate a full-sized image of a shoe impression—would be able to hold some 19.5 megapixels. Remember that the negative has about 14 megapixels equivalent, so the small print is limited by the size of the print that is being made and the larger one is limited by the detail on the negative itself.

Table 17.1 and Figure 17.5 summarize the findings. In Table 17.1, it is clear that the film images fell right into place with the digital cameras when the N = 3 criterion was used. Figure 17.5 shows what should be expected from various technologies and formats. It clearly shows that the medium format film should give decidedly better results than the digital cameras and the 35 mm camera, which is exactly what was found in the footwear study. This is especially true when a print is made. The prints shown in the chart are from the 120 film.

The bottom line is that, in practice, 35 mm film is typically:

- No better than a 14-megapixel camera for scanned negatives
- No better than an 8-megapixel camera when making prints

Table 17.1 *Results of Testing*

			Actual Frame	Critical Frame Widths					
Camera	**P width**	**D obj**	**Width**	**5 L/in**		**10 L/in**		**20 L/in**	
Canon				230		115		58	
Rebel XT	3456	108	85	0.37	Y	0.74	Y	1.47	N
		216	153	0.67	Y	1.33	N	2.64	N
Megapixels	7.96	324	200	0.87	Y	1.74	N	3.45	N
Canon				205		102		51	
10-D	3072	126	57	0.28	Y	0.56	Y	1.12	N
		252	110	0.54	Y	1.08	Y	2.16	N
Megapixels	6.29	378	147	0.72	Y	1.44	N	2.88	N
		504	196	0.96	Y	1.92	N	3.84	N
Kodak				309		135		69	
T-Max 100	4273	67	47	0.15	Y	0.35	Y	0.68	Y
		119	84	0.27	Y	0.62	Y	1.22	N
Megapixels	13.75	218	156	0.50	Y	1.16	N	2.26	N
		334	239	0.77	Y	1.77	N	3.46	N
		562	403	1.30	N	2.99	N	5.84	N
Print									
Limit due to 7.8 L/mm and 25% linear lens less									
4 × 6, B&W	3566	67			Y		Y		Y
Megapixels	9.54								

The table shows that when the frame width is critical or better (shown in green), the appropriate bars are resolvable. When the frame width was greater than critical, the bars were not resolvable (shown in red). The one minor exception was for the Canon 10-D; at 252-inch object distance the ratio was 1.08 instead of 1.00 and the bars were still distinguishable. Of importance, when the N = 3 condition was applied to the film images, they fell right into place.

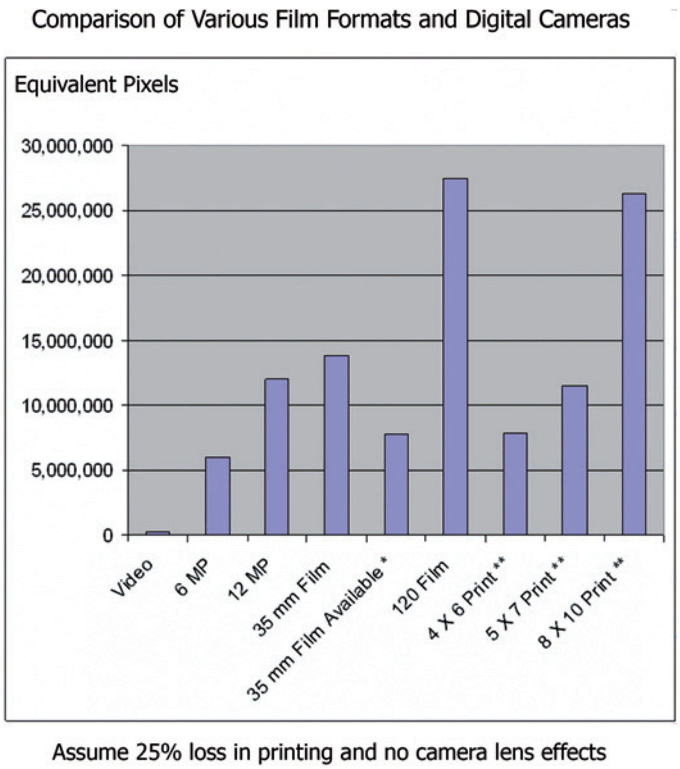


FIGURE 17.5 *Comparison of Film Formats and Digital Cameras. The bars indicate the number of equivalent pixels for each scenario. Compared to the 120 film printed onto an 8 × 10, all the others are rather similar to each other. This helps explain why in the footwear study the examiner did not really see any differences between 6 MP, 12 MP, and 35 mm. But they clearly saw the effect of the 120 film.*

DYNAMIC RANGE

The *dynamic range* is the light input range over which the sensor records fine increments, not output bit depth as described in Chapter 2. At the high-brightness end, the ability to respond is limited by saturation of the sensor elements. At the lower end, the ability to respond is clouded in random noise.

To measure dynamic range, several areas must be sampled *in a single frame*. Each area is at a specified level of brightness. We then measure the response of the light sensor and plot response versus input brightness—usually this axis is the horizontal axis and it is a logarithmic scale. The range of input brightness levels must be greater than the expected response of the system under test.

We then determine the point at which the system starts to respond, and the point at which it can no longer respond, and read off the indicated light levels. In order to make the exposure, we use a *sensitometer*. This device produces a reasonably uniform column of light and passes that column through a test target. Note that a transmission target must be used since any reflection target would be limited by front surface reflection. The camera under test takes a picture of the test target, and the camera output is measured. In the case of film, a camera is not necessarily needed. The film is simply placed in contact with and on top of the test target. The test target is simply a series of patches that absorb increasing amounts of the incoming light. The target is calibrated, usually in density units, since these correspond to log exposure to the sensor.

In the case of film, the density of the processed film is measured as an indicator of response; and in digital cameras, the images brightness levels are recorded. For measuring dynamic range, the units are not overly important since we are seeking the points at which the response curve approaches the horizontal asymptotes. Film manufacturers post the response curves of their main products on their web sites; extrapolating from that information, we can easily measure the dynamic range for each product. Digital camera manufacturers do not provide this data, so it must be measured. Measurements for three digital cameras are shown in Figure 17.6. Figure 17.7 shows a digital

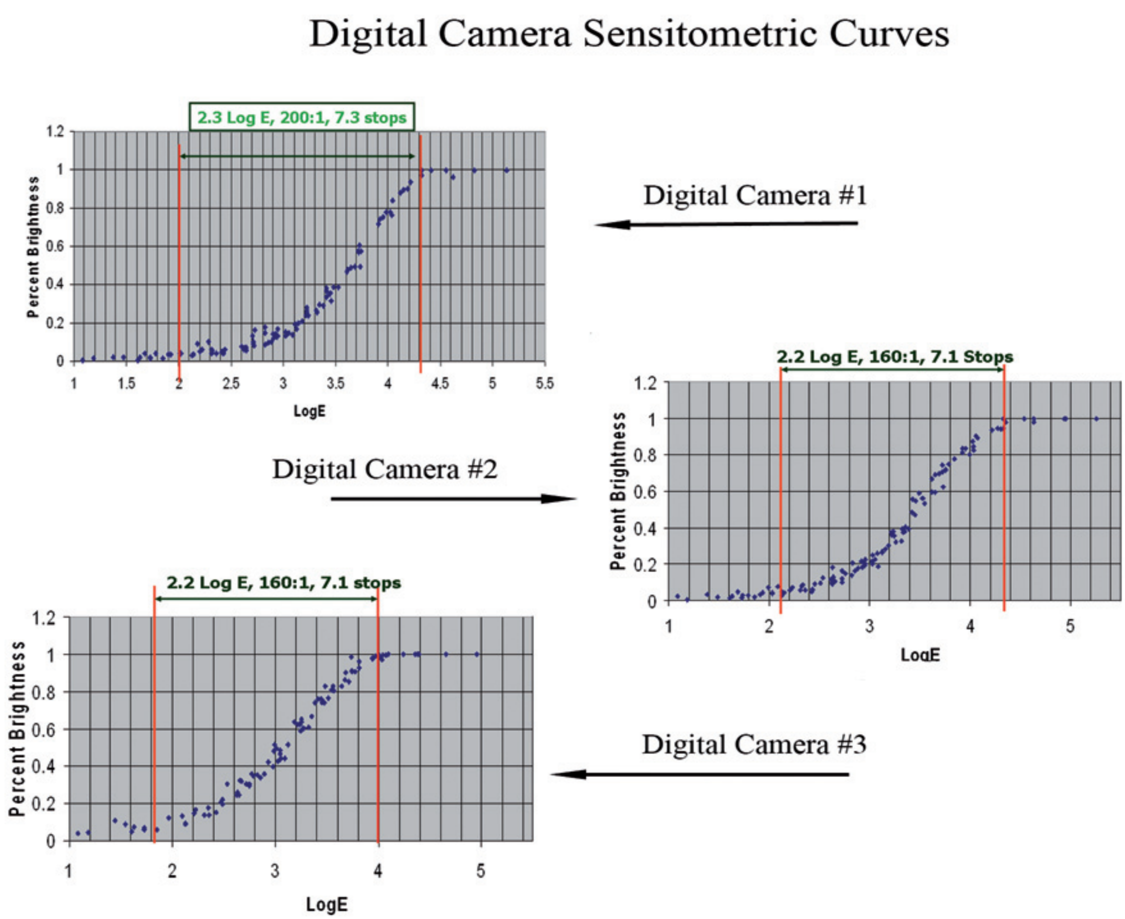


FIGURE 17.6 *Digital Camera Sensitometry. Sensitometric curves for three digital cameras are shown and the dynamic range is marked off for each. They are all very similar to each other and on the order of 2.2 Log E or 200:1.*

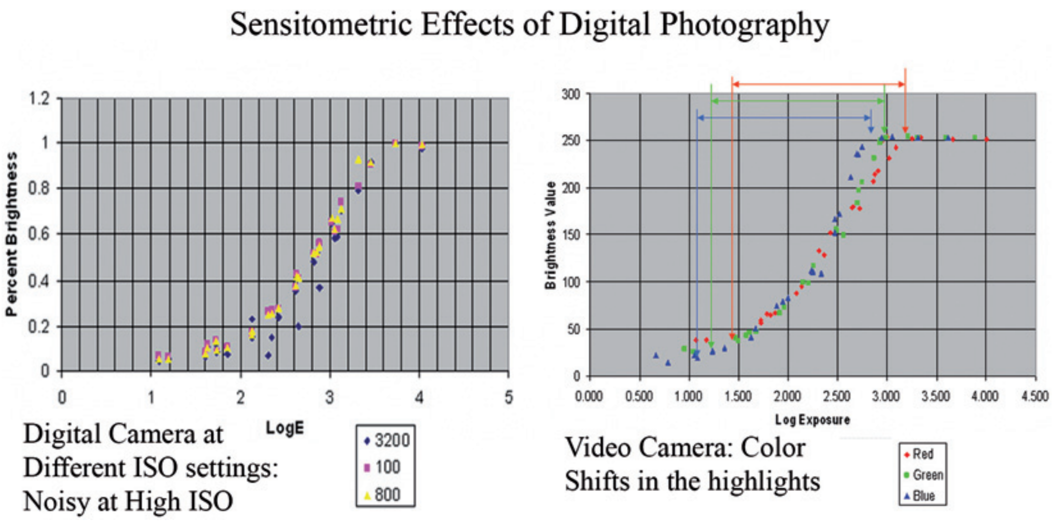


FIGURE 17.7 *Digital Camera Sensitometric Special Effects. The plot on the left shows that when the ISO on the camera was set to its highest level, there was a considerable amount of noise. The plot on the right shows a typical video camera, with a dynamic range of less than 100:1.*

camera set to a high ISO; the increase in noise is obvious. It also shows the extremely limited dynamic range of a typical video camera. Figure 17.8 shows data from the indicated web sites for certain film products. In summary, film cameras, using negative films, have the ability to record at least a 10,000-to-one range, and digital cameras are more restricted to a range of a few hundred-to-one. Slide films are closer to digital cameras.

What this means is that, when taking photos of scenes with a wide range of brightness levels, negative film will capture most of the information. With a digital camera, some information will not be recorded.

There are two solutions to this problem, but they apply only in certain circumstances. If the photo is of an inanimate object, as is the case in most forensic photography, one or both techniques can be considered. The simplest is to add fill light (see Chapter 9). This will brighten the dark areas. At crime scenes and accident scenes this might be hard to do, however. Another technique is to take several pictures at increasing exposure levels. The image contents will then be combined into a single image using software like the high dynamic range (HDR) tool in Adobe PhotoShop software. This tool takes the shadow information from the overexposed photos and the highlight information from the underexposed images and merges that into a final image that has a much greater effective dynamic range. This will work only if the subject is not moving. The effect is shown in Figure 17.9. There were originally five photos in this series, but only the second-most overexposed one and the second-most underexposed one are shown. These are the two on the left. Note that the detail in the upper left is not visible in the overexposed photo and the louvers on the device near the lower right are not visible in the underexposed photo. The photo on the right is the result

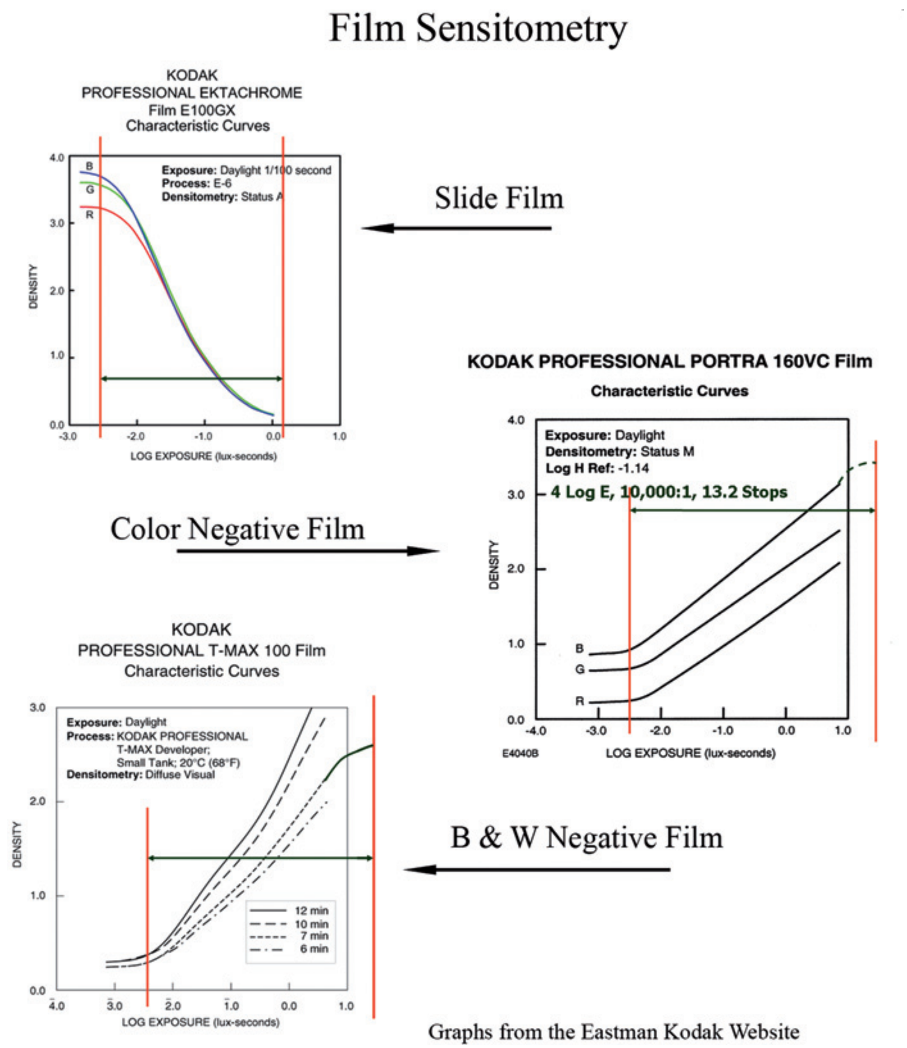


FIGURE 17.8 *Film Sensitometry. Sensitometric curves for three films are shown. The curves were downloaded from the Eastman Kodak Company web site. In some cases the bend at the top of the curve is not shown, so an estimated curve was put in place. Negative films have a dynamic range on the order of 4 log E, or 10,000 to one. Slide film is more comparable to the digital cameras.*

of merging the five photos in the series. Note that both the louvers and the upper-left detail are now rendered quite well. In situations where the subject is moving and there is no reasonable way to use a software tool or add fill light, use negative film.

SUMMARY

Putting the factors together, it becomes apparent that today's digital photography tools and techniques are very comparable to 35 mm film photography.

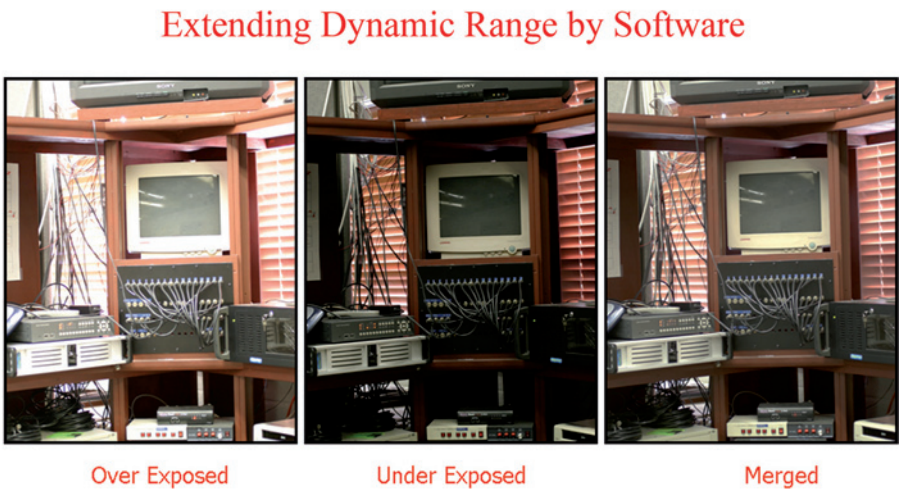


FIGURE 17.9 *Extending Dynamic Range by Software. Five images of the same subject were taken at different exposure levels. The two extremes are shown on the left. These five were then merged using the High Dynamic Range tool in Adobe Photoshop software suite. The result is shown on the right.*

One of the most important factors is resolution, and it is clear that 35 mm film photography is about the same as a 14-megapixel camera. And though most of the digital cameras today have less than that, a growing number of commercial offerings are approaching that value. All serious contenders today are within at least half that value. Larger format film cameras (for example, 120-size) will greatly exceed digital cameras; but if it is possible to take a few digital photos, systematically displaced from each other, it is possible to stitch those images together and rival even larger-sized film. Of course the subject has to be stationary for that to work, but that is often the case in forensic photography. Gazing into the crystal ball, it is doubtful that commercial cameras will go much above 14 megapixels. Not because of a technical limitation so much as because there is little need to do so. The increase in resolution will only be visible in very special circumstances, which most photographers will not encounter.

The second important factor is dynamic range. Currently film technology allows a range of at least 10,000-to-one. Digital cameras have a range of only few hundred-to-one. There are many scenes with elements ranging in brightness by more than that, and it is important to capture those elements. If the digital photographer can bring in fill light, that will solve the problem. If fill light is not an option, then there are postprocessing tools that can help if the subject is not moving. So the unresolved problem situation is a moving subject where fill lighting cannot be used. A new chip design from the Eastman Kodak Company, however, may bring a solution to this. It will result in higher chip sensitivity and probably increased dynamic range, as well.

In both film and digital photography, it is important to make sure that the print-making process is up to the quality demands of the photographic assignments. It is easy to lose much of the resolution and the dynamic range with a poor choice of printing technology.

CHAPTER 18

SWGIT

The Scientific Working Group on Imaging Technology, SWGIT (pronounced “swig it”) is an organization comprised of people from various agencies in the United States and a few other countries. It develops guidelines for the use of imaging technology in legal matters. It is one of several such groups (called “swigs”) that are sponsored by the United States Federal Bureau of Investigation. It was formed in 1997 and publishes documents containing guidelines. These are posted on the web site of the International Association for Identification (www.theiai.org). Since most of the association’s output is publicly available and kept reasonably up to date, that material will not be duplicated here. Instead we will focus on clarifying some of the key elements of the SWGIT guidelines.

PRIMARY AND ORIGINAL IMAGES

With film photography, the immediate result of snapping the shutter in the camera is the creation of a latent image. This is very ephemeral and is not visible until the film is processed. Once processed, there is a visible and physical image, the negative or slide. The negative or slide is the “original” image. But what happened to the latent image? It is gone—not thrown away, but merely converted to a more durable and usable form. Something very similar happens with digital photography. The snap of the shutter creates an ephemeral and invisible image on a memory card. Like a latent image, it is not durable, is easily altered, and cannot really be preserved for future reference. Like the latent image, it is converted to a durable and more visible form. Typically the contents of the digital file are converted to a more permanent form. It might be duplicated onto a CD, or copied to a suitable server’s hard drive (per Chapter 15). In some way it is preserved in a form that is independent of the camera and the memory card. The image that was on the memory card is called a primary image and the first recording of it onto a more durable medium is the original image. Primary images, be they latent film images or direct-from-the-camera digital images, cannot be kept because of the nature of the technology. Instead they are carefully and accurately preserved in a way that allows them to serve as future references. Since the film latent image and the negative or slide are on the same physical entity, it is possible to call the negative or slide both the primary and the original image. Primary images are treated with care, duplicated, and then discarded. Original images are treated with care and kept. Any subsequent image processing is done of duplicates made from the original.

IMAGE FILE AND IMAGE CONTENT

The value of a crime scene photo is that it can show what was in a particular place at a particular time and how the items were juxtaposed. That information is the content of the image. An image of a fingerprint is valuable because it shows the patterning of the ridges, pores, and other features, and it allows for measurement and analysis. This information can be extracted and the analyses performed on numerous presentations (either on screen or on a print) made from the same original even if they were made with slightly differing image parameters.

Consider the following experiment. You put film in your camera and take a picture of a bloody screwdriver lying on a table. You then take some photos of other items on the table and come back around to the screwdriver. You take a second picture, ostensibly from the same spot and using the same illumination. Later, after the film is processed and prints are made, you compare the photos. They are very similar, but by no means are they exactly the same. Nonetheless, the size, shape, blood pattern, manufacturer's label, and the nick in the top of the tip are easily discernable. The images are not identical photos, but they convey the same information content. Which is the right image? They both are! When prints are made from a negative, either the printer automatically makes adjustments or the operator makes them manually. Typically these are to balance the color and adjust the overall density. Never are precisely the same adjustments made twice. So, if multiple prints are made from the same negative, especially if other prints are made in between the repeats, which of the multiple prints is the right one? They all are since the analysis of the content will be the same even if some of the technical details are different print to print. Remember that the dyes used in color negative film fade a tiny amount every time bright light is shone on them—that is to say, the dyes fade a bit with each print. Thus the multiple prints cannot be exact duplicates of each other. Again, they are all correct because the content analysis remains unchanged by the technical differences in the images.

In the field of computer science, exact duplication is crucial for certain projects. This means that the original and the result of the duplication process are bit-for-bit exactly the same. There may sometimes be discussion about where the file starts and ends. There is the content of the file, its header, its trailer, its juxtaposition among other files on the same drive, and so on. But in the case in question, when a *duplicate* is made it is bit-for-bit exactly the same. If it is not, then it is a *copy*. For example if the file is a Microsoft Word document, and it is saved as a Word Perfect document, it cannot be a duplicate. Anyone reading the document will not be aware of the differences in file type, and the analysis of the content will be the same. But the analysis will have been made from a copy and not a duplicate.

In the case of imaging, the same holds true. For example, if we take a 16-bit RAW file, and make a 16-bit viewable TIFF file from it, it will be a copy and not a duplicate. It is virtually impossible to imagine a circumstance under which the content of the image will be any different but the TIFF is a copy and not a duplicate (also remember that RAW image files are not directly viewable). If the issue at hand had to do with the source of the original image or how and when it was actually recorded, there will be a difference. But as to the analysis of the content, there will be no difference. So, although the pristine nature of a duplicate may have some appeal to the naïve, it is not a substantive issue unless the provenance of the image is the issue, as opposed to its content.

THE MYTH OF PERFECTION

Although SWGIT does not address this overtly, it is inherent in the published documents. Some examiners, appealing to the idea of best evidence, will argue that the purest form of the image from the camera is the one that is the pure and perfect image. And since the least processed image you can take from a camera is the RAW file, they will argue that only this is the best evidence. This is not true.

The RAW file is not a viewable image. It is only a string of unprocessed data from the pixels in the camera. In order to view the image, it has to be converted to some sort of nonraw format. There are many computations involved in this and all of them involve some sort of estimation. The image finally viewed is the result of the original data as processed by algorithms that involve estimation (e.g., the demosaic) and use coefficients that also require estimation (e.g., color temperature). The viewable image is not nearly as pristine as the naïve might assume. Moreover, if the RAW file is converted to a viewable image using software from a third party as opposed to that provided by the camera manufacturer, the number and magnitude of the estimations is increased. There is no getting around the fact that the overwhelming majority of digital cameras fundamentally undersample the image and the images they produce must include means to artificially correct for the undersampling. There is no perfect image! And, for other reasons, storing RAW files in pursuit of perfection is not a good idea.

Camera manufacturers provide software to convert RAW files into images, but they tend not to support those modules for more than three or four years. Meanwhile, in the forensic community, it is often important for images to be kept for decades. Adobe Corporation provides a module that can be used for most of the major cameras, and they claim they will maintain that for a very long time. Without either the manufacturer's software or the Adobe software it is impossible to convert RAW files into images. The Adobe software, though quite good, contains certain compromises that are made to accommodate a wide range of cameras. The camera manufacturers put significant effort into the mathematics inherent in their conversion software and believe that this gives a performance advantage. They are not willing to share this knowledge. The result is that when the investigator gets around to converting stored RAW files decades after the picture was taken, he or she will have to use software containing compromises if he or she can get it at all. Had the images been converted to viewable form and stored in an open format such as TIFF, these compromises are eliminated. The penalty is that TIFF images, since they contain all three color channels for all pixels, are some three times as big as RAW files, which have only one channel for each pixel.

Another important issue arises from the question of the settings to be used in converting the images from RAW to viewable. The camera will make some estimates of what the settings should be. These can be applied inside the camera automatically to make very high quality JPEG images, or someone can process each and every image and use the camera suggested settings. If the settings are to be changed, there is the question of who is doing the conversions. If it is the photographer, then he or she can set them according to what he or she saw at the scene. If it is someone else in a different city or a different decade, then the changes are merely guesses. In light of this, what was the advantage to saving RAW files?

There is frequent argument regarding how many colors a person can actually discern. The generally accepted maximum value is about 16 million. Conveniently this is achieved by having 8 bits per channel for three channels. All printers and projectors have no more than 8 bits per channel. So even if you have more than 8 bits per channel, the bulk of that information will be eliminated when the image is shown to people. RAW files record 16 bits per channel, because of all the arithmetic that is performed on the binary, integer numbers will cause round-off errors. This is especially important in converting the native linear analog brightness values captured by each sensor element into a logarithmically scaled digital number. But once this is done, it is very hard to argue that the content of an image will be interpreted differently if it is viewed in 16 bits as opposed to 8. In fact it is not possible to view it in 16 bits. So, which is perfection: the unviewable RAW file, the RAW file converted to a viewable image involving compromises, the RAW file converted to a 16-bit TIFF file soon after the photos are taken, or a RAW file converted to an 8-bit TIFF image soon after the pictures are taken? The message is that there is no perfection in either film or digital photography and taking extreme measures in an attempt to preserve it is an effort taken in vain.

TYPE I AND TYPE II IMAGES

Almost all the legal contests involving the use of digital images have dealt with how the images were processed and who did the processing. At the same time, the bulk of the images, digital or otherwise, receive very little additional processing at all. And, that which was applied was very simple and easy to understand. Most images are used to show generally what something looked like and where items were in relation to each other. The only image processing that is normally applied in these instances might involve a bit of color and overall brightness adjustment. These images are referred to as Type I images. They could almost be replaced by drawings. They are described by a witness and it is the testimony of the witness that will get the scrutiny—not the veracity of the images. With Type I images the investigator probably needs to indicate only what steps were employed to prepare the image for presentation without going into any detail.

By comparison, some images are subjected to extensive analysis. The investigator might use an FFT technique to reduce interference from background patterns. Or a set of images might be brought to specific sizes in order to make comparisons. Tone scales might be distorted in order to show artifacts, and colors might be drastically changed to better show certain features. There are far too many possibilities to enumerate them here. These images are referred to as Type II images. Lay audiences (including judges and juries) will need good explanations regarding how the original image was processed in order to arrive at the one being shown. They have to be convinced that the image being shown is a legitimate extension of the information contained in the original. Also, in a legal proceeding, the opposing counsel may bring in their own expert to contest the validity of the processing employed. Accordingly it is prudent to keep good notes of the steps employed: which tools, what sequence, and what settings were used.

In summary, the investigator should be prepared to indicate the steps taken to prepare Type I images in a general sense. With Type II images he or she should be prepared to explain in great detail the nature of the processing steps and explain why those were legitimate.

IMAGE PROCESSING TASKS

There are three main groupings of activities that are involved in the processing of images. These are referred to as *tasks*. They are not job descriptions since in different organizations they might be done by any of several different people or all by the same person. They do have different skill set requirements, however.

The first task is *image preparation*. This involves activities such as downloading images from the medium taken from the camera and the application of the basic steps required to identify the image, protect and store it, and make it ready for basic viewing and potential analysis. This may involve the creation of a working image and the use of basic tools such as color balancing and brightness adjustment. The operator needs to be fully trained in how to perform these basic steps. If basic image enhancements are applied, the image preparer should record the steps employed. Strict adherence to a standard operating procedure is required. In most cases, this person will not be called to testify, and if he is, it will be related to the provenance of the image and the fact that he did not do any analysis or anything that might endanger the integrity of the image. This task is more commonly done by a separate individual when traditional film is used. Please note that in the case of film, the color balancing and brightness adjustment changes are made automatically by the printing machine. The record of these steps will not be known, but the process involved *is* known.

The second task is *image enhancement*. This involves the application of analytical tools in order to bring out information that is not readily apparent from viewing the basic image. There are literally thousands of potential processes that can be applied and selection of the approach will depend somewhat on knowledge of how the content will be interpreted. For example, if a latent fingerprint is to be enhanced, the operator should know basically what to look for in determining the detail of a print so as to show best what is required, *and* at the same time not hide features that might change the assessment.

In some cases more detailed knowledge of a particular discipline is required, but in others this is less of a requirement. For example, if the issue is to determine whether the lock button on a door is engaged or not, there is no specific forensic discipline and most people familiar with such locks are able to determine whether there is a good rendition of the lock button. In some cases the forensic discipline is extremely complex, for example forensic medicine or dentistry. In these cases, the person doing the image enhancement either should be the content expert, or should work collaboratively with that expert. The person doing image enhancement will probably be called to testify as to what was done and why it was done that way. Significant knowledge of image processing tools and techniques is required. In the case of film photography, this task often is done by a photography expert and not necessarily by the content expert. It is also true that the range of enhancement tools normally applied in film photography is not as extensive as the range commonly used in digital work.

The third task is *image analysis*. The steps involved in image analysis are not so much involved with the application of image processing tools as with the application of expert judgment regarding the information extracted from the image. For example the person determining whether a given latent fingerprint matches a given exemplar or not requires detailed knowledge of fingerprint examination science, tools, and techniques. The person doing this should be a certified fingerprint examiner and will definitely be called upon to testify. Likewise, a bitemark should be analyzed by a forensic odontologist, and bodily wounds should be analyzed by a forensic pathologist.

In the earlier paragraphs of this chapter, the idea was that it was the conclusion drawn from the content that is paramount in forensic imaging work and not the exact parameters of the image itself. This comes to the fore in the image analysis task. If the image is processed in two different ways and one of them leads to the person doing image analysis to draw a bad conclusion, then that image processing was faulty. It is for this reason that the image enhancement task might be done by the content expert or at least in collaboration with him or her. Sometimes it is prudent to take a few different approaches to verify that the conclusion drawn is robust.

It should be noted that the odds of a global, general purpose image processing step including a particular individual by mistake are infinitesimal. It is much more likely that a null conclusion will be found. The typical image has some six million pixels and each can assume any of some 16 million values. The probability that a specific cluster of these will be altered in just such a way as to identify a specific, uninvolved person is unimaginably small. By comparison, the odds that an image will prove inconclusive are much higher.

IMAGE ENHANCEMENT

There is a virtually endless list of ways in which an image can be enhanced. With Type I images the objective is to make sure that the images have a realistic appearance that closely emulates that of the original scene. With Type II images the objective is to render certain details in the image more visible and analyzable to the content expert. The analyzed image may look very different from the original image or the original scene, for that matter. If a forensic pathologist uses some form of false colorization in order to more clearly see and possibly measure a patterned injury, the resulting image will look very different from both the original image and the original scene. But the process can be quite valid for its intended use.

Some tools are used simply to make the image appear more natural. These include color balance, brightness adjustment, contrast adjustment, and unsharp masking. Other tools are used more analytically; for example, resampling and resizing an image to facilitate superimposition of two or more images. Frame averaging to reduce random noise or highlight defective pixels is another more advanced tool. The Fourier transformation can be used to eliminate unwanted and distracting background patters. Wiener filtering can be used in an attempt to remove fundamental problems such as motion blur and out of focus conditions. SWGIT provides a long document on this topic.

STANDARD OPERATING PROCEDURES

Forensic image analysis has recently been brought into the fold of disciplines that the American Society of Crime Laboratory Directors (ASCLD) recognizes.

It is grouped with computer forensics, forensic video analysis, and is labeled as Digital and Multimedia. Likewise, the American Academy of Forensic Science has created a new section called Digital and Multimedia Science. Forensic photography, which is the taking of the pictures, is not directly included. Instead, the person drawing conclusions is expected to assure that the photography was done correctly. In order for a laboratory to pass an inspection by the ASCLD Laboratory Accreditation Board (ASCLD/LAB) it must have Standard Operating Procedures for each area being examined. The ASCLD/LAB mantra is, "Write what you do and do what you write". The "write what you do" part essentially says you need Standard Operating Procedures (SOPs).

Forensic laboratories are, after all, scientific laboratories and they need to assure that their work complies with valid scientific procedures. One way to do this is to hire only the top experts in each field and assume that they know all there is to know about their fields. Or, they can take the combined wisdom of many experts and build guidelines for their personnel in the various fields of study. In the former approach each investigator is fully and personally responsible for each and every step they take; in the latter, many of the detailed issues already have been worked out by groups of experts, and lab investigators can work on the basis of their assessments. Clearly the latter makes the most sense.

Following the approach of assessment of key issues by a group of experts, the SWGIT has met for some 10 years and over time has developed guidelines for many of the issues associated with the use of imaging in forensic applications. It has published several documents that offer guidelines for developing laboratory-specific SOPs. Those raise and discuss the issues, indicate the crucial points, and offer examples of SOPs. SOPs should not be overly detailed or bogged down in step-by-step instructions, but they should contain the key steps that will lead to scientifically valid investigations and reports. Much of the step-by-step information is included in the training that laboratory personnel take, and this will change as new equipment and software is brought into use.

AUTHENTICATION, INTEGRITY, AND ARCHIVING

The topics of authentication and integrity are often lumped in with archiving, and the grouping may be called by any one of the three words. They are related but different. They are stated in a nominally sequential order but need to be considered separately. The SWGIT documents include separate documents on these three topics.

Authentication deals with the matter of whether an image is a fair and accurate representation of what it purports to show. This starts with how the photograph was taken in the first place. Many of the famous photos from the U.S. Civil War are not authentic—they were staged.

At that time, photographic plates used in cameras were wet coatings of material on glass plates. The photographer had to make the solutions, coat the plates, and then make the exposures before the whole thing dripped away. A photographer could not simply point and shoot. Action shots were a rare occurrence. So the photographers would lay out the bodies in poses that they thought showed the story they wanted to tell. Then they would prepare the plates, make the exposures, and process the images. The images are not fair and accurate of the reality, but rather were depictions of the story telling imagination of the photographer.

In one recent news story, a photograph was published showing many refugees standing dejectedly behind a fence. Another photo of the general scene showed that the fence was really just a fragment of a fence, about 30 feet long and there was no fencing on either end. The people just happened to be standing by the fence and selectively framing the photo through the fragment of fence provided a nice newsworthy image—engaging, yes; fair and accurate, no. So-called "snuff films" purport to show people being murdered. Are they authentic? Often the forensic video analyst is brought in to try and determine whether a real crime has been committed.

There are several tools that can be used to try and determine whether the content of an image is a fair and accurate depiction or a misleading one. But, it is a difficult task and not always successful. The key issue is to recognize that there is no image filing software product that will address the true substance of authentication. At trial, the photographer, or someone else there at the time, will testify as to whether the image being shown is fair and accurate.

Integrity deals with whether an image has been altered in some way that will degrade the ability of viewers to correctly ascertain the state of the object that the image shows. Altering an image to mislead an audience must be prevented, either purposefully or accidentally. If an image is to be enhanced, a working copy of the original will be made and the working copy will be enhanced. In this way the original remains as a reference starting point. The enhanced image becomes a new image.

There are two key elements to integrity: protection and demonstration. *Protection* involves steps taken to assure that access is limited to those authorized to have such access and that inadvertent damage is not likely during storage. These steps might include both physical and logical measures such as password or stronger limits on access to servers, keeping records in an area open only to authorized personnel, and keeping back-ups in an off-site location.

Demonstration deals with the ability to prove that the image that was stored away is not changed during storage. Two of the more common approaches are the use of visual verification and hashing. Visual verification involves the person storing the images away viewing them at that time and then, when retrieving them, viewing them again to assure that the image shows the same basic content. Hashing involves the application of a hash algorithm to the original image before it is stored away, and then checking the hash number when retrieving the image.

Archiving is the process of storing an image for an extended period of time and assuring that the images that were stored away are retrievable when needed at some time in the future. Different jurisdictions have different requirements, but most are multidecade or longer requirement for evidentiary materials. The key issue with archiving is that it is an active process and not a passive one. We do not put images into a cave that will last forever and then simply go into the cave to carry them out—although there are such arrangements in old abandoned salt mines for documents. This not viable for digital files because, unlike plain paper documents, there may not be a way to read the electronic files in the distant future.

The difference between digital files and paper documents is that in order to use a digital file, you need a compatible computer, reading device, and software. With paper documents all you need is the ability to find them; then they are immediately readable. Very few laboratories have the ability to read punched paper tape any more. And even if they had the devices, they may not have the software drivers for the devices that are compatible with today's operating systems. And finally, even if they had the device and the driver they would probably not have the application program that wrote the original file. The result is that the punched paper tape has become a write-only memory. This sequence is not isolated but rather quite common in the rapidly evolving world of digital technology.

Since there is an ongoing march of technology, the archivist must refresh records as technology turns over. Records on old media must be transcribed as newer media emerges and as new readers and writers replace the current technology. Files must be transcribed to new formats as old formats or the software required to read and write them vanishes or becomes incompatible with more current counterparts. In summary, maintaining an archive for digital records is not a passive activity, but a careful and active one.

IMAGE QUALITY

The first thing that most people think of when they consider the quality of digital images relative to film images is image quality. The most salient aspect that they think of is that of resolution. In a pure technical measurement sense, mid-2000's cameras are very close to having the resolution of film negatives. Another aspect of image quality that might come to mind is dynamic range. Most mid-2000's cameras have a dynamic range comparable to that of slide film and much less than that of negative films (see Chapter 19). Other aspects you might consider, but that don't immediately jump to mind are tone scale (digital cameras have a sharp cut-off in the highlights portion of the range) and color quality (digital cameras have gotten quite good). But all of this can be a false path. The real issue is analysis of the content of the image.

An image needs to be as good as the users require in order to draw sound conclusions from the photos provided. Even if the film image in a particular discipline is technically better than that of a digital camera, if the difference does not alter the conclusions drawn by the subject matter expert, there is no effective quality difference. This expands to the photographers' techniques. The lower dynamic range of digital cameras can be compensated by balancing the light when the photo is taken. The lower resolution can be compensated by working with a smaller object frame size or by taking two photos and stitching them together. The sharp cut-off in the toe can be compensated by slightly underexposing. And for dynamic range and sharp cut-off issues, if the image is of a still subject, we can take multiple photos and then combine them with dynamic range expanding software tools.

In summary, the normally applied judgments regarding the quality of digital photography to film photography can be misleading. The real issues center on the requirements of the subject matter expert drawing conclusions, the tools and skills of the person doing the image enhancements, and the tools and techniques of the photographer. As a practical matter there are very few situations where 35 mm film photography is necessarily better than high-quality digital photography. Larger format film is a different matter and one that has not yet been studied in the forensic field.

VIDEO IMAGING

SWGIT devotes considerable effort to the forensic analysis of video, but that is not the subject of this book. If video processing tools are used to capture a specific frame as a still image then all of the material in this book is applicable to that frame.

CHAPTER 19

Digital Images and the Investigative Process

The prior chapters have given you a solid understanding of the technical "magic" behind digital images and how they're taken, created, produced, enhanced, altered, and analyzed. Now it's time we looked at the nontechnical and legal issues surrounding their credibility, reliability, and use as evidence inside a legal court of law. You would think that all the details, time, man-hours, and effort exerted into preparing the digital image for trial would be the hardest part of your job. However, you will find that preparation for court, including organizing all your documentation and procedures along with your expert testimony, may very well be the toughest hurdle to overcome.

Although you should leave the legal issues up to the legal team and counsel on the case, you will need to understand these issues in order to effectively assist counsel with use of the digital evidence. Preparing your evidence will be your first major hurdle, getting your digital evidence admitted into court will be your second, and keeping it there will be your last. All computer and digital media examinations are different. The examiner must consider the totality of the circumstances as he or she proceeds. Therefore, not all components here may be needed in every situation, and examiners may need to adjust to unusual or unexpected conditions in the field. Just remember to document what you do, especially any deviations away from your normal procedure.

Evidence preservation and integrity have been discussed elsewhere in this book, but the important part will fall under your Standard Operating Procedures (SOPs): how you preserved your evidence, maintained the integrity of the evidence, and documented your actions. You must be able to defend all aspects of the digital evidence, from how the evidence was preserved to the sound integrity of the evidence. Your documentation must reflect how well you followed your SOP, methodologies, and must assist in defending your actions—everything that was done to the evidence.

To prepare for getting your digital evidence into court we will spend some time discussing documentation, SOPs, Chain of Custody (COC), the Federal Rules of Evidence, and legal precedent. This chapter will be devoted to standard operating procedures, chain-of-custody, and forensic procedures and the next chapter will discuss the legal rules and related case law.

STANDARD OPERATING PROCEDURES

Standard operating procedures are a good place to begin. SOPs are policies, procedures, and practices used by the examiner and/or agency for any given tasks and situations. They should contain detailed information on your tasks, methodologies, workflow, report formats, approval process, and any other activity done in the course of your analysis or exam.

SOPs should be the lifeblood of an expert/examiner, right alongside their forensic tools. Forensic examiners should be following and closely adhering to the set of SOPs established in their laboratories in order to be consistent and have a methodology that they know is valid and reliable. Additionally, SOPs can show that examiners have followed and documented regular procedures. SOPs assist the examiner to achieve uniformity, consistency, and reliability at any given tasks and allows for peers to duplicate the task to verify consistent results. For the examiner, not following your SOPs leads to the same results as not having any SOPs, or even worse. The American Society of Crime Laboratory Directors, which accredits crime labs, has a motto, "Write what you do and do what you write." Without SOPs, the expert runs the risk of noncompliance, poor reproducibility, inconsistent results, and potentially rendering the evidence (results of the analysis and investigation) inadmissible in court. There are two options: 1) explain your actions by explaining the science, mathematics, and technology you used to a lay jury, or 2) indicate that you followed SOPs developed according to industry-accepted guidelines.

Therefore it is important to develop "defensible" SOPs—defensible because they need to hold up to the tough scrutiny of the court in a legal proceeding. In fact, everything the expert does and uses needs to be defensible. You should be able to defend and support your SOPs, methodologies, tools, equipment, and actions. Use your SOPs religiously and use them to help you explain (or defend) any deviations or variations from those methodologies or procedures.

There are no standards to date for forensic operating procedures. However, there are several resources available to assist in developing a defensible set of SOPs. The National Institute of Justice, through the Department of Justice, has several resources available. The Scientific Working Group on Imaging Technology (SWGIT) provides guidelines that can help the practitioner conform to forensic requirements. Check out the list of references and resources at the end of the book for additional information.

The following are some ideas that you should consider including in your SOPs:

- Establish a methodology: a set of working principles, postulates, practices, procedures, and rules used by those who work in a discipline or engage in an inquiry
- Develop a peer review process using qualified examiners
- Use processes that are repeatable
- Be consistent in your use of a process: do it the same way every time
- Document exceptions and deviations, including reasons
- Verify routinely that equipment is working properly
- Document that equipment is routinely checked and validated
- Require continuing education and training for all investigator/examiners
- Require proficiency testing (examiners and operators of equipment and tools)
- Validate and test tools routinely
- Require written documentation of analysis, actions, and conclusions
- Make sure in reports that the evidence validates and supports the conclusions

Keep in mind that forensic examinations are rarely similar. They are unique and different—no two are alike. As the examiner runs into unexpected or unusual circumstances, which happens more often than not, the examiner must adjust and document accordingly. He or she must consider the totality of the circumstances, including field conditions, scope of exam, tools, and so on. Keep good documentation so the work done by the examiner can be reproduced by another examiner to validate the results, if necessary. Being able to point to a set of SOPs used for the analysis will be additional weight that the court will consider when determining the value of the results.

The use of a defensible SOP can be valuable when getting your evidence admitted into court. In a 2007 federal court Daubert hearing (Daubert will be discussed in the next chapter), the prosecutor was trying to get the testimony of a government examiner regarding digital evidence in a criminal case—*U.S. v. Nobumochi Furukawa, U.S. District Court, District of Minnesota, Criminal Case Number 06-145 (DSD/AJB)*—admitted into the court trial, but the defense moved to exclude the testimony. The government examiner referenced and identified Section 14 of SWGIT as an indication of community acceptance of image/video authentication and the judge ruled that the testimony did meet Daubert and was relevant, so it was admitted.

This use of SWGIT guidelines in this case is an excellent example of how and why those SOPs are so very important to the examiner and getting the digital evidence into court. If it isn't admitted into court, it cannot be used to assist in the case. There are a number of ways the SWGIT guidelines work with the legal rules of evidence and it is instructive to understand how they work together.

AUTHENTICATION ISSUES

There are a number of points provided that can help establish that images are authentic. First of all, with camera images, the SWGIT guidelines call for duplicating images, in an unopened form from the camera's removable medium onto a medium that is more useful to long-term storage and preferably create an uneditable record. This is referred to as creating an original image from a primary image. In addition, the guidelines address the use of camera heading information and hashing algorithms to help people later demonstrate that reasonable caution was taken to preserve the integrity of the images. The use of chain of custody procedures is also recommended. With video, the media presented to the operator should be evaluated to help assure that it is an original, and steps should be taken to prevent changes to the original. SWGIT goes on to define Type 1 and Type 2 images. Type 1 images are intended for showing where things are in a scene and not intended for in-depth analysis. They can be visually verified, which is to say the witness can state that, "I was there, and this photo is a fair and accurate representation of what I saw." Type 2 images are ones that will probably be subjected to significant processing and analysis. This means that the resulting image may not look very much like the original scene, even though it is derived from it. In this case visual verification may not be feasible and extra precautions must be used. Transmission of images should be sufficiently secured to assure that the integrity of the image is not corrupted or accessible to those without permission to have the images.

VALIDITY AND RELIABILITY OF SCIENCE EMPLOYED

In many instances, lawyers lump validity and reliability together and call the result reliability. To the scientific community, "validity refers to the degree to which a measuring technique measures what it purports to measure" and "reliability refers to the extent to which a measuring instrument produces the same result when it is used repeatedly." (Modern Scientific Evidence, Vol 2, Faigman, Kaye, Saks & Sanders, West Books, 1997). SWGIT takes a separate path and considers the scientific notion of validity to be reliability and the scientific notion of reliability to be repeatability.

Nomenclature not withstanding, the person intending to use scientific tools to prepare images for use in court must be prepared to show that the underlying science is valid and reliable. This is the essence of the typical pretrial hearing under Rule 702/Daubert or Frye in other states. Accordingly, SWGIT indicates that all image processing software used should be well known in the industry or you should be prepared to speak to its validity and reliability. When equipment and procedures outside the guidelines of SWGIT are used, the operator should be prepared to explain and defend the processes employed. This is especially true for Type 2 images, where a log of the steps taken should be kept. SWGIT indicates that the documentation should be such that a trained operator can take the original and the listing of steps taken and produce an image that would lead to the same conclusion as that of the original operator.

Further, certain image processing tools should be identified as ones that should not be used unless there are special circumstances, and the operator is prepared to defend his or her actions. Lossless compression is preferred, but lossy compression can be used at low compression ratios (almost lossless). If unusual or ad-hoc analytical procedures are employed, the operator should be prepared to demonstrate that these approaches are valid and reliable by citing testing of the tools, or use by other practitioners in relevant situations.

PROPER APPLICATION AND INTERPRETATION

In order to show that the procedures were consistent with the underlying science and that the conclusion is a clear result of the raw evidence, use the SWGIT procedures and pay attention to points of certain caution. First of all, the agency should have SOPs and make sure that they are followed. Equipment should be regularly calibrated and tested. Personnel should be trained for their assignments, tested for competency and regularly tested for proficiency. Image capture, processing, and reproduction equipment should be demonstrably adequate for the assignment. And, there should be a list of relevant reference materials to help operators as they do their work. For Type 2 images, a log of the steps taken should be kept, and where particular accuracy is required, proper notes of steps taken should be recorded. SWGIT has several documents that give detailed guidelines for certain types of photographic and imaging assignments.

THE CONCLUSION DRAWN IS PROPER

Using images to support testimony may require that three types of tasks are performed, and witnesses generally should not testify to tasks they did not perform, are not qualified to perform, or do not have firsthand knowledge of what was done. SWGIT defines these tasks as follows:

- Technical Preparation: These tasks include authentication, basic image enhancement, and preparation of displays and images for additional analysis.

- Examination: These steps are used to determine properties of the image itself, including determination if the image is computer generated, whether it has been processed and or watermarked, and so on.
- Interpretation: Analysis of the subject matter of the image to draw conclusions about that subject image's content. This requires expertise in the domain of the content.

Agency management should be advised of the major image evidence issues and assure that the handling of such material is done according to well-developed SOPs, and that the operators have the proper training and tools.

CHAIN OF CUSTODY

Chain of custody (COC) procedures should be well documented and incorporated into your standard operating procedures depending on laboratory policies. A COC document should memorialize the complete journey of evidence during the life of the case—from the moment it is acquired to the very end of its life. It should identify the who, what, when, where, how, and why with respect to the handling the evidence. Whenever the evidence is moved, transferred, or any action is taken upon the evidence, it should be thoroughly documented. Once physical custody has been taken of the evidence, it is essential to initiate and maintain a COC that properly documents that the evidence was handled in a manner that will insure it can be properly identified, authenticated, and admitted into court or other legal proceeding. A COC form is designed to maintain control and accountability of the evidence from acquisition to disposition of the evidence item. Although some jurisdictions do not call for a formal COC for certain images because of interactions with other local laws and regulations, the basic concept should be employed to help assure integrity.

A COC document should, at a minimum, collect the following information regarding a specific evidence item:

- When, how, why (purpose) evidence is acquired, transferred, moved, or stored
- Full description of the evidence; identification marks
- From whom it was acquired
- Who handled that evidence and their signature
- To whom it was transferred; who took possession of the evidence and their signature
- Who took it out of storage; – when and why

- The name of the transferring entity should match the previous "transferred to" entity
- What procedures or action was used on the evidence
- Where the evidence is located at each step/stage; account for all storage of the evidence
- Date and time evidence was acquired, transferred, moved, or stored
- If an overnight delivery courier is needed to transfer the evidence, attach the courier's signature documentation to the COC (Case law regarding mailing evidence and signature requirements is well documented. Usually, the mailing of evidence does not, in and of itself, break the chain of custody; see *Pasadena Research Laboratories v. United States*, 169 F.2d 375, 380-81 (9th Cir.1948), cert. denied, 335 U.S. 853, 69 S.Ct. 83, 93 L.Ed. 401 (1948); *Rosedale Coal Co. v. U.S. Bureau of Mines*, 247 F.2d 299 (4th Cir.1957); *Gallego v. United States*, 276 F.2d 914 (9th Cir.1960).)

The COC documentation should accompany the evidence and be maintained at all times, since it may be needed as part of the authentication process for the evidence in a legal proceeding. Without a clearly documented COC, or if there is any break in the sequence of documented actions on the evidence, the evidence may lack needed credibility, rendering it very difficult, if not impossible, to be admitted as evidence in a legal proceeding. Physical evidence is admissible when the possibilities of misidentification or alteration are "eliminated, not absolutely, but as a matter of reasonable probability." *United States v. McFadden*, 458 F.2d 440, 441 (6th Cir.1972). "Challenges to the chain of custody go the weight of the evidence, not its admissibility." *United States v. Levy*, 904 F.2d 1026, 1030 (6th Cir.1990), cert. denied, 498 U.S. 1091, 111 S.Ct. 974, 112 L.Ed.2d 1060 (1991). Remember, the jury can weigh the importance of testimony and evidence based upon the credibility they assign to the witness—and the opposing counsel will help them to apply very little credibility if he or she can.

Therefore, if you start your chain of custody with the acquisition of the digital evidence, properly mark and identify the evidence, track and maintain all the required signatures, secure storage, and inventory logs through the disposition of the evidence, the COC should remain intact and valid. Keep in mind that the opposing side in a legal battle is looking for any way to keep damaging evidence out of court. What better way to prevent good evidence from being introduced into court, than to attack and show a break in the chain of custody, raising doubt as to the authenticity of the evidence in question.

CHAPTER 20

Getting Digital Images Admitted as Evidence at Trial

Whether digital or enhanced digital photographs are admitted into a legal proceeding or court as evidence is determined by a collection of evidence rules. Courts also interpret these rules based on the particular issues and facts in each case, which can make it more than a little difficult to prepare a clear-cut checklist to follow for getting your evidence admitted in a legal proceeding. The Federal Rules of Evidence (FRE), used by the federal court system, apply to Electronically Stored Information (ESI) or computerized data just as they apply to other types of evidence. (MANUAL FOR COMPLEX LITIGATION § 11.447, 4th ed., 2004). Currently there are no admissibility requirements set forth in the Federal Rules of Evidence specifically for digital photographs. In a Georgia Supreme Court Case on October 5, 2001, *Almond v. State*, 553 S.E.2d 803 (Ga. 2001), the court stated that digital pictures were properly admitted after the foundation was laid that would be the same for any other photograph. "We are aware of no authority, and appellant cites none, for the proposition that the procedure for admitting pictures should be any different when they were taken by a digital camera." Therefore, at least in Georgia, traditional issues of authentication and relevancy will govern.

In the recent court case of *Lorraine v. Markel American Insurance Co.*, 2007 U.S. Dist. LEXIS 33020 (D. Md. May 4, 2007), Judge Paul W. Grimm proposed that whenever electronic data (this would include digital photographs) is offered as evidence, the following evidence rules should be considered:

1. Is the ESI *relevant* as determined by Rule 401 (does it have any tendency to make some fact that is of consequence to the litigation more or less probable than it otherwise would be);
2. If relevant under 401, is it *authentic* as required by Rule 901(a) (can the proponent show that the ESI is what it purports to be);
3. If the ESI is offered for its substantive truth, is it *hearsay* as defined by Rule 801, and if so, is it covered by an applicable exception (Rules 803, 804 and 807);

4 Is the form of the ESI that is being offered as evidence an *original* or *duplicate* under the original writing rule, or if not, is there admissible secondary evidence to prove the content of the ESI (Rules 1001-1008); and

5 Is the probative value of the ESI substantially outweighed by the danger of **unfair prejudice** or one of the other factors identified by Rule 403, such that it should be excluded despite its relevance. Preliminarily, the process by which the admissibility of ESI is determined is governed by Rule 104, which addresses the relationship between the judge and the jury with regard to preliminary fact finding associated with the admissibility of evidence. Because Rule 104 governs the very process of determining admissibility of ESI, it must be considered first.

Not all digital photographs will have to go through all these steps, but be assured that they may have to go through some, all, or additional scrutiny by the courts, depending on the individual case. Therefore the hurdles to overcome in getting digital evidence admitted into federal court could include Federal Rules of Evidence 104, 401, 403, 901, 801, 1001–1008. If the digital evidence for admissibility is photographs or digitally enhanced photographs, Federal Rules of Evidence 701 through 705 will probably be added to the preceding list of rules. The Rules of Evidence are complex and not easy to understand, except by attorneys and legal professionals, and they will be discussed and explained more fully later in this chapter. Also, the key elements of the actual Federal Rules of Evidence can be found in the appendix.

For years, photographs have been authenticated under Rule 901 by the testimony of a witness who could state that he was familiar with the scene and testified that the photograph was a fair and accurate representation of the scene. Digital photographs offer a new set of authentication issues for the court, since they can be more easily manipulated, altered, or enhanced. As you will note in the case discussions that follow, authentication of digitally enhanced photographs requires an explanation of the process and methodology used to enhance the digital photographs by someone knowledgeable about the enhancement process and tools.

In the *Lorraine v. Markel* case, Judge Grimm wrote a 101-page opinion of the admissibility of electronically stored information, which, of course, includes digital photographs. For a more in-depth tutorial or primer on the admissibility of electronic evidence, read Judge Grimm's 101-page opinion (see *Lorraine v. Markel American Insurance Co*., 2007 U.S. Dist. LEXIS 33020 D. Md. May 4, 2007, available on the web at http://www.mdd.uscourts.gov/Opinions/Opinions/Lorraine%20v.%20Markel%20-%20ESIADMISSIBILITY%20OPINION.pdf).

The Federal Rules of Evidence govern the rules of evidence for federal courts and the state courts have their own rules of evidence. However, most state courts have accepted some form, if not all, of the Federal Rules of Evidence for their state courts. It is always best to check with the court system where your case is being litigated for their rules specific to that court.

The following is a brief overview of some of the relevant Federal Rules, as amended in 2000, available at http://judiciary.house.gov/media/pdfs/printers/109th/evid2005.pdf or http://www.law.cornell.edu/rules/fre/index.html.

RULE 104. PRELIMINARY QUESTIONS

a **General questions of admissibility.**
Preliminary questions concerning the qualification of a person to be a witness, the existence of a privilege, or the admissibility of evidence shall be determined by the court, subject to the provisions of subdivision (b). In making its determination it is not bound by the rules of evidence except those with respect to privileges.

b **Relevancy conditioned on fact.**
When the relevancy of evidence depends upon the fulfillment of a condition of fact, the court shall admit it upon, or subject to, the introduction of evidence sufficient to support a finding of the fulfillment of the condition.

c **Hearing of jury.**
Hearings on the admissibility of confessions shall in all cases be conducted out of the hearing of the jury. Hearings on other preliminary matters shall be so conducted when the interests of justice require, or when an accused is a witness and so requests.

d **Testimony by accused.**
The accused does not, by testifying upon a preliminary matter, become subject to cross-examination as to other issues in the case.

e **Weight and credibility.**
This rule does not limit the right of a party to introduce before the jury evidence relevant to weight or credibility.

The rules of evidence are administered by the court trial judge. Rule 104 of the Federal Rules of Evidence requires that the court determine preliminary questions concerning the admissibility of evidence, qualification of a person to be a witness, or the assertion of privilege. If relevancy depends upon facts such as authentication or personal knowledge—a condition of fact—then Rule 104(b) requires a strong showing of the existence of the condition as a prerequisite to admitting the evidence. In other words, there is evidence sufficient to support a finding by a reasonable juror that it is more probable than not that the condition is true.

RULE 403. EXCLUSION OF RELEVANT EVIDENCE ON GROUNDS OF PREJUDICE, CONFUSION, OR WASTE OF TIME

> *Although relevant, evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, or by considerations of undue delay, waste of time, or needless presentation of cumulative evidence.*

This rule is known as the P & P rule, which deals with whether the evidence is more probative than prejudicial, probative tending to prove or actually proving a particular proposition or to assist the trier of fact (the jury, or, depending on the case, the judge) of the truth of an allegation. Prejudicial refers to the degree to which the testimony and evidence elicits the tendency for preconceived judgment or convictions that are not elemental to the case. If the conclusion drawn by the witness claims a finding to be so certain that it is virtually unchallengeable, then the conclusion will impinge on the duty of the finder of fact to establish just what is fact and what is probable. FRE 403 authorizes the judge to exclude evidence if its probative value is "substantially outweighed" by the potential of unfair prejudice. (United *States v. Plaza*, 179 F. Supp 2d 444, E.D. Pa. 2001). "Relevant evidence" means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence—FRE 401. Relevancy doesn't require certainty of the fact being offered, only some tendency that makes the fact more probable. This rule allows admitted testimony to be challenged as to weight. This means that the trier of fact can determine the degree to which they believe what was presented and how much it bears on the ultimate determination of guilt. Using our adversarial system, opposing counsel can always challenge the weight.

RULE 702. TESTIMONY BY EXPERTS

> *If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.*

In other words, the expert must be able to demonstrate all of the following:

- The expert has sufficient knowledge, experience, skill, training, or education about the subject on which he is testifying.
- The testimony is based upon sufficient facts or data.

- The testimony is the product of reliable principles and methods (i.e., Standard Operating Procedures).
- The expert applied the principles and methods in this particular case (i.e., expert followed the Standard Operating Procedures, which are based upon tested analytical tools and procedures).

The standard for qualification of an expert witness is a liberal one. The test to be applied when qualifying an expert witness is whether the witness has any reasonable pretension to specialized knowledge on the subject under investigation (*Commonwealth v. Wallace*, 817 A.2d 485 (Pa.Super.2002)). If the witness has the specialized knowledge, he or she may testify and the weight to be given to such testimony is for the trier of fact to determine. *Id.* Accordingly, an expert witness does not need formal education related to his or her area of expertise, and he or she may be qualified based on training and experience alone. That is, the court may allow someone with limited qualifications to serve as an expert, but the opposing counsel can be expected to make it clear to the jury that the expertise is, in fact, questionable.

The court will look at the examiner's background and skills, when qualifying an expert. So an impressive background, experience, and CV will help present a solid background. Having impressive credentials always helps add to your credibility, but isn't mandatory. You want the court to rule that the expert is qualified as an expert under Rule 702 because his or her "scientific, technical or other specialized knowledge will assist the [judge] to understand evidence or determine a fact at issue." Additionally, be professional in your dress, speech, and demeanor. As an expert, you will be scrutinized by how you look as much as how you talk. Equally important are the expert's communication skills. An expert must be able to present the facts in clear and concise terms that are not too technical and "above" the heads of the judge or jury, yet technical enough to be informational and factual. An expert must fully know and understand his or her procedures, findings, technical processes, and documentation well enough to withstand detailed scrutiny and probing questions by the opposing counsel. The expert may be competing against the opposing side's expert and be cross-examined by the opposing counsel, leaving the trier of fact (judge or jury) to decide who is more believable. The testimony by the expert must appear to be knowledgeable, fair, neutral, and objective.

DAUBERT AND FRYE

This is a good place to discuss scientific evidence and the use of the Frye or Daubert test. Up until 1993, the Frye test was used for admissibility of scientific evidence in federal courts. The Frye test was set forth in *Frye v. United States*, 293 F. 1013 (D.C.Cir.'23), where the court said that admissible scientific evidence must be grounded in a theory that had "general acceptance" in the relevant scientific community. The judges in *Frye* ruled that:

> *Just when a scientific principle or discovery crosses the line between experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.*

Where novel scientific evidence is at issue, a Frye inquiry allows the judiciary to defer to scientific expertise precisely as to whether or not it has gained general acceptance in the relevant field and will exclude any evidence from test results that arrive out of unproven scientific methods. The trial court's gatekeeper role in this respect is conservative, thus helping to keep "pseudoscience" or junk science out of the courtroom. The Frye standard was then adopted by the California Supreme Court in *People v. Kelly* (1976) 17 Cal.3d 24; you will sometimes hear this called the Kelly-Frye standard.

The Daubert test came from the 1993 US Supreme Court case of *Daubert v Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993). It requires four things to be shown:

- Whether the theory will help fact-finder (that is, whether the theory has or can be tested)
- Whether the theory has been peer reviewed (this is a component of good science and will improve the likelihood the flaws will be found)
- Whether the theory has a significant rate of error
- Whether the theory is "generally accepted" (not required, but it helps to know that the scientific community approves, and to demonstrate this by citing the literature)

Federal Rule 702 is the result of admissibility issues captured in the Daubert criteria and codifies the criteria from three cases: Daubert, Joiner, and Kumho Tire—now called the Daubert Trilogy (see *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993); *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999); and *General Electric Co. v. Joiner*, 522 U.S. 136 (1997)). In the states that follow Daubert, the rule closely resembles the federal rule. In Kelly-Frye states, it is more like the findings attributed to the two named cases. Frye applies to novel usage, truly scientific processes, and relies on the scientific community for guidance as to what is accepted and what is not. Daubert leaves the admissibility decision more to the court, and although it alludes to the matters of science and process, it does not rely so heavily on the scientific community. It also can apply to nonnovel applications and to technical issues as well as scientific ones. There is no definitive count for the number of states relying on Daubert or Frye, since State courts are split on following Frye or Daubert. Some follow a variation of either, depending on the facts in the case.

Rule 702 is referred to as the R & R rule. It deals with the reliability (scientifically, the validity of the science and the repeatability of the technique) and the *relevancy* of the testimony—that is, the degree to which it applies to a proper element in the case that is in dispute and which helps the trier of fact resolve that issue and decide the case.

RULE 703. BASES OF OPINION TESTIMONY BY EXPERTS

> *The facts or data in the particular case upon which an expert bases an opinion or inference may be those perceived by or made known to the expert at or before the hearing. If of a type reasonably relied upon by experts in the particular field in forming opinions or inferences upon the subject, the facts or data need not be admissible in evidence in order for the opinion or inference to be admitted. Facts or data that are otherwise inadmissible shall not be disclosed to the jury by the proponent of the opinion or inference unless the court determines that their probative value in assisting the jury to evaluate the expert's opinion substantially outweighs their prejudicial effect.*

As an example for this rule, an expert need not enter into the record proof that the speed of light is about 300,000 kilometers per second. An expert may testify to information other than that presented at trial, if that is of a nature that is widely held in the relevant community or can reasonably be assumed for the case at hand. Information that is otherwise inadmissible cannot be disclosed to the jury unless its probative value outweighs its risk of prejudice of the jury misunderstanding in its need to assist the jury to evaluate the expert's testimony. Rule 703 does not prohibit, and therefore invites opposing counsel to cross-examine the expert on his opinion, which may open the door to additional underlying facts or evidence that might not otherwise get introduced—our adversarial system at work.

RULE 704. OPINION ON ULTIMATE ISSUE

a Except as provided in subdivision (b), testimony in the form of an opinion or inference otherwise admissible is not objectionable because it embraces an ultimate issue to be decided by the trier of fact.

b No expert witness testifying with respect to the mental state or condition of a defendant in a criminal case may state an opinion or

> inference as to whether the defendant did or did not have the mental state or condition constituting an element of the crime charged or of a defense thereto. Such ultimate issues are matters for the trier of fact alone.

The opinions drawn by the expert witness may not state the guilt or innocence of the defendant, or the truth or falsity of allegations or matters of the law. These are reserved for other parties.

As an example, the expert can state that it is his or her opinion that this is the defendant's fingerprint, but he or she cannot say that the defendant is therefore guilty of the crime.

Many forensic specialties have stock phrases that they use to avoid problems with this rule. For example a medical examiner might state that his or her conclusion has reasonable medical certainty, or a hairs and fibers expert may say that the characteristics are consistent. DNA experts will state the probability of the sample belonging to another person. SWGIT has declined to prepare such statements, but each jurisdiction can prepare a set of stock phrases to use.

RULE 705. DISCLOSURE OF FACTS OR DATA UNDERLYING EXPERT OPINION

> *The expert may testify in terms of opinion or inference and give reasons therefore without first testifying to the underlying facts or data, unless the court requires otherwise. The expert may in any event be required to disclose the underlying facts or data on cross-examination.*

The witness may testify to opinion without specifying underlying data unless the court requires otherwise. Rule 705 governs the disclosure of the facts underlying expert testimony and allows experts to testify to opinions with or without disclosing the underlying facts. Rule 705 remains silent on the inadmissible part of an expert's basis or the evidentiary status of the inadmissible part of an expert's basis.

> *Rule 705, which permits experts to give opinions without previously disclosing their underlying factual basis, is inconsistent with the screening role assigned to the presiding judge, inconsistent with the independent fact-finding role of the jury, and generally incompetent litigation strategy.* (Federal Rules Of Evidence Advisory Committee: A Short History of Too Little Consequence, *191 F. R.D. 678 (2000), Paul R. Rice and Neals-Erik William Delker)*

When the expert does not disclose the basis for his or her opinions on direct examination, any worthy opposing counsel will pursue the underlying data and facts from the expert during cross-examination.

RULE 901. REQUIREMENT OF AUTHENTICATION OR IDENTIFICATION

- a General provision. The requirement of authentication or identification as a condition precedent to admissibility is satisfied by evidence sufficient to support a finding that the matter in question is what its proponent claims.
- b Illustrations. By way of illustration only, and not by way of limitation, the following are examples of authentication or identification conforming with the requirements of this rule:
  1 Testimony of witness with knowledge.…

The full text of this rule can be found in the Appendix at the end of this book.

The preceding Federal Rules of Evidence, when applied in conjunction with properly documented guidelines or SOPs, such as the documents from the Scientific Working Group on Image Technology (SWGIT), can assist in the authentication, reliability, and validity of digital image evidence.

EVIDENCE

Evidence can take many different shapes or forms from direct to circumstantial, and the types of evidence are many. According to the Sixth Edition of Black's Law Dictionary, evidence is "any species of proof, or probative matter, legally presented at the trial of an issue, by the act of the parties and through the medium of witnesses, records, documents, exhibits, concrete objects, etc., for the purpose of inducing belief in the minds of the court or jury as to their contention… Testimony, writings, or material objects offered in proof of an alleged fact or proposition." You might say there are three main categories of evidence: (1) real, (2) demonstrative, (3) testimony. Real evidence is physical evidence that is relevant to the case, such as a weapon found at the scene of a murder, or a contract in a contract dispute. Demonstrative evidence is visual aids or materials used to illustrate or assist the court in understanding some issue in the case, such as simulations, demonstrations, or recreating an event or thing relevant to the case. Testimony is defined, strictly, as utterances that are meant to be believed on the teller's say-so alone, not because of supporting arguments or any like considerations. Demonstrative evidence is also used to help explain or illustrate oral testimony. Digital photographs, especially enhanced digital photographs, will normally fall into the demonstrative evidence category as they are used to help explain or illustrate.

Precedent is very important in helping lawyers, experts, defendants, and all players in litigation matters understand what they're up against in "winning" their given case or how to construct their strategy. Precedent, is defined by Harvard Law School's online dictionary as: "A judicial decision that a court should follow when deciding a similar case. To serve as a precedent for a current case, a prior decision must have a similar legal issue and a similar factual scenario. Our common law system is based upon precedent as courts seek to decide current cases on the basis of legal principles established in prior similar or analogous cases." (see http://www.law.harvard.edu/library/services/research/guides/united_states/basics/one_l_dictionary.php).

Looking at cases involving digital evidence, enhanced images, and other related issues that have been resolved in the courts is both very instructive and the best way to evaluate how courts may behave in the future given specific issues. These cases, or precedent, serve as the basis for future decisions even more than laws and statutes themselves, since the precedent shows how the courts have interpreted the laws in the given jurisdiction. It is extremely important that the laws, statutes, and precedent in each state and jurisdiction be examined. Although most states pattern their rules of evidence and procedure after their federal counterpart, they each have their own variations and differences. Additionally, if working outside the United States, you will need to check with each country's jurisdiction and laws for how they handle any particular issue.

There are several ways to research case law in your jurisdiction from using a local law library to searching the Internet. Most attorneys use a database from a paid service provider such as Westlaw, Lexis, LoisLaw, or CaseMaker, but the Internet offers several ways for the nonattorney to get fairly reliable information on cases. There's always Google, but The National Clearinghouse for Science, Technology, and the Law (NCSTL) at Stetson University School of Law, offers a free database of forensic topics and relevant cases. The following are several recent cases depicting issues with digital evidence and digitally enhanced photographs used in court. You should find these case summaries instructive in developing an understanding of how the courts, in general, treat digital evidence and enhancement of images as evidence. Several of the following case summaries were taken from the NCSTL database and others through traditional search methods. It would be instructive to read the full case for additional information and to get the full impact of the issues and rulings.

State of Connecticut v. Alfred Swinton

A significant and often cited case regarding enhanced digital photographs is *State of Connecticut v. Alfred Swinton, 268 Conn. 781 (2004); 847 A.2d 921*, available at http://www.jud.state.ct.us/external/supapp/Cases/AROcr/CR268/268cr60.pdf.

Relevant Case Summary

Defendant was convicted of murder. The evidence against defendant included bite marks on the victim, identified by a forensic odontologist, as those belonging to the defendant "with reasonable medical certainty." On appeal, the defendant contested the admissibility of the bite mark evidence, photographic overlays, specifically: *"[1] photographs of a bite mark on the victim's body that were enhanced* using a computer software program known as Lucis, and [2] images of the defendant's teeth overlaid, or superimposed, upon photographs of the bite mark that were made through the use of Adobe Photoshop, another computer software program." The state claimed that the enhancement photographs were simply a way to present the evidence to the jurors, much like a PowerPoint demonstration. The defense argued that the images were altered by the enhancement process and the reliability of the process needed to be established.

After a lengthy and complex discussion on the enhanced photographs of the bite marks, the court found that the State had laid a proper foundation for the enhanced photographs by having a witness who was well-versed in the program used to create the enhancement. The court denied the defendant's claim. However, the court decided that the witness who testified about the computer-generated dentition overlays on bite-mark photos lacked the computer expertise to satisfactorily explain the image before the jury. As a result, the court held that the overlays were improperly admitted. Regardless, the Court held that the admission of the overlays was harmless and affirmed the judgment of the trail court.

New criteria for laying a proper foundation for the introduction of computer-generated evidence was established by the Connecticut Supreme court in this case. The issue was whether the evidence offered by the State satisfied the requirements of the confrontation clause—from the Sixth Amendment in the United States Constitution which provides:

> *In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed, which district shall have been previously ascertained by law, and to be informed of the nature and cause of the accusation,* ***to be confronted with the witnesses against him****, to have compulsory process for obtaining witnesses in his favor, and to have the assistance of counsel for his defense.*

The confrontation clause protects the rights of the accused to "confront those witnesses against him" through cross-examination of witnesses and is heavily used by opposing counsel for cross examination of expert witnesses.

At trial, the State introduced photographic evidence showing bite marks on the victim. An Adobe Photoshop superimposition of the Appellant's teeth over the bite mark was introduced through a forensic odontologist who did not produce the image and had little understanding of the program. The State also introduced a set of computer-enhanced photographs created using the Lucis program through an experienced masters-level forensic scientist who produced the evidence himself and had knowledge of how the program functioned.

The Swinton court noted that one underlying theme in several cases is that the analyst or expert who testifies be the person who had engaged in the enhancement process so that he can testify to the specific detail of the process. See *English v. State*, 205 Ga. App. 599; *Nooner v. State*, 322 Ark. 104; *Dolan v. State*, 743 So. 2d 546; *American Oil Co. v. Valenti*, 179 Conn. 359. Swinton looked further into this issue by turning to the Federal Rule 901 (b) (9), which provides that authentication or identification of a process or system requires "evidence describing a process or system used to produce a result and showing that the process or system produces an accurate result."

The Court stated that the salient issue was reliability, not only of the materials themselves, but also of the process by which the materials were generated and the tools used in that process. The Court reiterated the rule articulated in *American Oil Co. v. Valenti*, 179 Conn. 349, 359, 426 A.2d 305, 310 (1979) (holding that there must be "testimony by a person with some degree of computer expertise, who has sufficient knowledge to be examined and cross-examined about the functioning of the computer") and adopted the six factors that establish authentication under Fed. R. Evid. 901.

The factors are:

> *1) the computer equipment is accepted in the field as standard and competent and was in good working order; 2) qualified computer operators were employed; 3) proper procedures were followed in connection with the input and output of information; 4) a reliable software program was utilized; 5) the equipment was programmed and operated correctly; and 6) the exhibit is properly identified as the output in question."* *(Swinton at 942)*

Commonwealth v. Serge, 837 A.2d 1255 (Pa. Super. 2003)

This case had a question of admissibility of computer-generated animation as demonstrative evidence. It was an appeal from the judgment of sentence entered in the Court of Common Pleas of Lackawanna County, Pennsylvania after a jury convicted Appellant of first degree murder for fatally shooting his wife.

Relevant Case Summary

The court determined that to admit computer-generated evidence a party must "establish foundational requirements such as authentication, relevance, fairness and accuracy in representing evidence, and probative value exceeding possible prejudice." Normally, evidence in court falls into one of three categories: testimonial evidence, documentary evidence, and demonstrative evidence (see 2 McCormick on Evidence § 212; John W. Strong et al. eds., 5th ed.,1999). The purpose of demonstrative evidence is to help the trier of fact understand the other evidence. A trial court may admit demonstrative evidence, just like any other evidence, if the relevance outweighs any potential prejudicial effect (*Commonwealth v. Reid,* 571 Pa.1, 811A.2d 530; 2002).

The court in *Commonwealth v. Hatcher*, 746A.2 1142 (Pa. Super. 2000) reasoned that noninflammatory photographs are admissible if relevant and helpful to the jury's understanding of facts, and inflammatory photos are also admissible if their probative value outweighs potential prejudicial effect. The Reid court noted that "Demonstrative evidence, however, must also be properly authenticated by evidence sufficient to show that it is a fair and accurate representation of what it is purported to depict. Pa.R.E. 901(a). Demonstrative evidence may be authenticated by testimony from a witness who has knowledge of what the evidence is proclaimed to be. Pa.R.E. 901(b)(1)."

The court found that the admission of the computer-generated animation was permitted given that the Commonwealth satisfied all foundational requirements for admitting the evidence under Pa.R.E.901 as follows:

1. A forensic expert authenticated the animation by describing the process behind the animation, explaining how the animation was created, and stating that it was a "strict depiction" and accurate representation of the Commonwealth's expert opinion.
2. The trooper involved and an expert witness on the shooting confirmed that the computer animation was a fair and accurate depiction.
3. The computer animation was a "clear, concise, and accurate depiction of the Commonwealth's theory of the case," was relevant under Pa.R.E. 401, and aided the jury in understanding the numerous testimonies.
4. There was no unfair prejudice under Pa.R.E. 403 to the appellant from admission of the computer animation, since it offered a "uniquely vivid and cohesive rendition of the collective testimonies about the shooting," supplying probative value to the animation.
5. The computer animation contained no "inflammatory content," in that it did not offer audio sound or detailed graphics of "facial expressions, evocative movements, or evidence of injury such as blood." Thus complying with the pretrial directive that the animation be devoid of drama to prevent any prejudicial effect.
6. The trail court gave the jury an appropriate cautionary instruction sufficient to guard against confusion.

In 2004, the Pennsylvania Supreme Court granted allowance of appeal in this case to consider only the admissibility of the computer-generated animation and made its final decision on this issue on April 25, 2006. (See [J-37-2005] In the Supreme Court of Pennsylvania Middle District, found at http://www.courts.state.pa.us/OpPosting/Supreme/out/J-37-2005mo.pdf.) The Pennsylvania Supreme Court determined that computer-generated animation (CGA) evidence is admissible as long as it is properly authenticated, it is relevant, and its probative value outweighs the danger of unfair prejudice or confusion.

CGA should be treated as any other demonstrative evidence. Thus it needs to satisfy the requirements for admissibility under Pa.R.E. 401, 402, 403, and 901. Accordingly, the Pennsylvania Supreme Court affirmed the decision of the Superior Court in the previous case.

RODD v. Raritan Radiologic Associates, P.A., 860 A.2d 1003, 373 N.J. Super. 154 (N.J. Super. Ct. App. Div., 2004)

Relevant Case Summary

Defendants appealed the judgment in a medical malpractice wrongful death action. At trial the plaintiff presented computer magnified images of digitally scanned selections of the 1997 and 1998 mammograms, magnified between 30 to 150 times the actual x-ray. The recognized diagnostic tool at this time was a handheld 2.5 power magnifying glass to render only four times magnification. The jury viewed these images projected onto a very large 6×8-foot screen. The defendant objected to the super-magnified images because no notice was given during discovery. [Discovery is the process that allows both sides in a legal action to acquire relevant "information" from the other party prior to trial in order to gain pertinent facts in the case. This is a pretrial process and is intended to give both sides the chance to gain sufficient information through depositions, interrogatories, documents, and other relevant forms of information to prepare their case for trial.]

The defense attorneys first learned of the existence of these images during the pretrial conference, too late to properly test the process for creating the images. Additionally the defendant objected to these super-magnified images due to the potential danger that the jury would not fully understand the potential for distortion and confusion that could result by the use of super-magnification. Halfway through the trial, the defense was provided with only the printout of selected images. The trail court admitted the images over objections and without any limiting instructions to the jury on their narrow purpose. In fact, the images were used extensively throughout the trial, suggesting to the jury that the magnified clusters were equally as clear on the mammogram film itself.

Noting that this type of evidence must be authenticated, relevant, and that its probative value must be weighed against any prejudicial effect, the appellate court found that the images were "unduly influential, may have created confusion, and could have been accepted as substantive evidence." Although the plaintiff countered that the super-magnified images were only for demonstrative purposes and not meant to be substantive, the jurors were never given instructions as to this limited purpose. The court found that the "visual aid" of these images "took on testimonial significance, and its contents were highly susceptible of being perceived by the jury as substantive evidence." The images were clearly capable of rendering an unjust result. The court was concerned with the lack of adequate notice given to the defense and the failure to lay a complete foundation for the introduction of these computer images. There was no account for how these super-magnified images were processed, created, or how the computer software operated. The court determined that the computer-generated exhibits were not supported by testimony of a witness with sufficient knowledge of how the exhibits were created and that a "computer-generated exhibit requires a more detailed foundation than that for just photographs or photo enlargements." The error in permitting these computer-generated exhibits was not harmless. The case was reversed and remanded for a new trial.

This case clearly shows the importance of understanding your tools, how and why they were used, and the need to be able to have sufficient knowledge to explain their use to a jury.

People v. Perez, No. D039428, 2003 WL 22683442 (Cal. Ct. App. 2003)

This case is not a "published" case. Cases not published normally are not considered precedent; that is, used as legal authority to support cases with similar issues or facts. However, this case should prove to be very instructive regarding the use of digital images in court.

Relevant Case Summary

This was a criminal case addressing the use of enhanced digital images of a shoe print. The defendant contended that a Kelly-Frye hearing should have been held prior to the admission of digitally enhanced shoe print evidence. The shoe print evidence was digitally enhanced using Adobe Photoshop software and was admitted at trial. The defendant contended that the digitally enhanced shoe print should not have been allowed as evidence because no Kelly-Frye hearing was held. Perez claimed that the Adobe Photoshop software used to enhance the image was a "new scientific" technique. The court, reasoning that the software "is not a scientific technique" but offers "an easier way" to develop film or pictures, ruled that the evidence was reliable since it did not alter the shoe print image.

State of Ohio v. Hartman, 93 Ohio St. 3d 274; 2001 Ohio 1580; 754 N.E.2d 1150 (Ohio 2001)

Relevant Rules

Pursuant to Ohio rules of evidence 702(B), a "witness is qualified as an expert by specialized knowledge, skill, experience, training, or education regarding the subject matter of the testimony." A witness does not need any special education or certification to gain expert status. The individual expert need only possess enough knowledge of the field in question, to aid the trier of fact in performing the fact-finding function. Ohio Rules of Evidence 104(A) provides that the question of admissibility of a person as an expert is determined by the trial court. "In making its determination it is not bound by the rules of evidence except those with respect to privileges."

Relevant Case Summary

Defendant was convicted of first degree murder, kidnapping, and tampering with evidence. Defendant appealed his conviction, raising 13 propositions of law. The Defendant contended that the trial court erred in admitting computer enhanced fingerprint images, since the State expert lacked the necessary qualifications to testify and that the digitally enhanced fingerprint evidence was unreliable.

At trial, the defendant claimed that the digital enhancement of the fingerprints was "blazing new ground" and challenged the scientific reliability of the method. However, the trial court concluded that: "the use of the computer in this instance is no different than… would be the use of an overhead projector, microscope, a magnifying glass or anything else like that that would enhance an expert's ability to make his determination and therefore I find that there's nothing—no new trails being blazed here and I'm overruling the objection for that reason." (Hartman, 754 N.W.2d at 1165)

The fingerprint expert testified that the fingerprint found at the crime scene was the defendant's fingerprint. The expert came to that conclusion by comparing the fingerprints using traditional methods and came to the same conclusion when he compared the prints to the digitally enhanced images. The defendant did not challenge the expert's credentials as a general latent fingerprint examiner nor did he challenge the fingerprint identification by traditional methods.

Therefore, after reviewing previous case law and the entire trial record, the appellate court affirmed the trial court's death sentence, holding that the evidence was properly admitted and the expert testimony from the State was admissible.

Bryant v. State of Florida, 810 So.2d 532 (Fla. Dist. Ct. App. 2002)

Relevant Case Summary

Defendant appeals his conviction of three counts of child abuse. Defendant argued that the trial court erred in admitting enhanced video time-lapsed videotape into evidence without granting a continuance to allow the defense to evaluate the original tape for comparison to authenticate. The only evidence that the jury was allowed to see regarding the alleged abuse was this edited and enhanced videotape. The defense asked for a continuance to compare this edited tape to the original to evaluate the state's expert's testimony regarding the alterations to the tape. Defense counsel objected that the edited, fully enhanced excerpt was inadmissible under the best evidence rule. The defense was deprived of the opportunity to consult an expert to assess differences between the original and edited, fully enhanced videotape. Of crucial importance, however, was the fact that tampering had occurred through the insertion of time lapse sequences to create the real-time effect of standard videotapes, which could more likely persuade the jury of the defendant's guilt.

The appellate court concluded that admitting the edited enhanced version of the videotape over the defense's objection for a continuance for comparing the excerpt with the originals was in error. The case was reversed and remanded for a new trial.

Kennedy v. State of Florida, 853 So.2d 571 (Fla. Dist. Ct. App. 2003)

Relevant Case Summary

The defendant was convicted of first degree murder and robbery and appealed. Defendant's sole contention on appeal was that the trial court erred in not holding a Frye hearing on the State's image enhancement techniques to determine if they were based on generally accepted scientific methods in the scientific community. At the crime scene, black and white photos were taken of bloody shoeprints. These were then digitally enhanced, using a Liquid Crystal Violet (LCV) to make the print more visible and changing the contrast to improve the clarity. Morhitz software was used to enhance the image prints. The expert testified that neither the LCV nor the enhancement process altered the evidence or created evidence. The expert on the software testified as to how it works and how it was used, explaining that the original image was never altered and that the software logged every adjustment done to the image.

Therefore the appellate court concluded that a Frye hearing was not required in that no new or novel scientific methods were used, and that the State did not alter or change the evidence. Therefore the trial court's conviction was affirmed.

State of Florida v. Victor Reyes (Fla. Cir. Ct. 2002)

This proceeding is the Defendant's motion *in limine* regarding latent prints. A motion *in limine* (Latin, meaning "at the threshold") is made before the start of a trial asking the court to rule that certain evidence may, or may not, be introduced to the jury in a trial. A motion *in limine* generally addresses issues that would be prejudicial for the jury to hear, even if the other party makes a timely objection, and the judge instructs the jury to disregard the evidence. This motion normally is heard in the judge's chambers and out of hearing of the jury. If the motion is granted, then evidence regarding the issue could not be mentioned in front of the jury, without first approaching the judge outside of the hearing of the jury and obtaining permission.

Relevant Case Summary

The defense requested the court to grant a motion *in limine* to prevent the admission of fingerprint evidence in an execution style murder case.

The defense claimed that the latent fingerprint captured on negative print film (from a print on duct tape) and then digitally enhanced using a software program named PC Pros MORE HITS was unreliable and untested. The defense claimed that the software was a "new and unverified science" that does not meet the Frye admissibility standard.

The full text of the opinion can be found at http://www.fdiai.org/FRYE%20hearing.htm, and would be extremely instructive to read, since it gives insight into how this court viewed the issue of digitally enhanced photos and the need for examiners to follow and use guidelines, such as SWGIT.

Interestingly, the State presented three highly qualified expert witnesses to explain the use of the software and the process of enhancement, whereas the defense offered only one expert witness who was an expert in Adobe Photoshop. The State's experts successfully explained the processes used, testified that they followed standard operating procedures and guidelines, explained that all enhancements were performed on exact copies of the original image, and that in the enhancement process "no areas of an image were deleted or altered in any way." One expert noted that the procedures followed "… the Scientific Working Group on Imaging Technologies' (SWGIT) recommendations and guidelines for the use of digital image processing in the criminal justice system. The stated purpose of the SWGIT guidelines is to ensure the successful introduction of forensic imagery as evidence in a court of law." Additionally, the court heard testimony that the MORE HITS program was used currently by several federal agencies as well as "more than 150 different state and municipal law enforcement agencies throughout the United States," Canada, and England.

The court found that the process of digital enhancement of videotapes, photographs, and fingerprints was widely used and accepted as reliable in the forensic community. This finding was based on the testimony of the experts and that the Automated Fingerprint Identification Systems (AFIS) has been successfully used for more than 30 years in the United States and other countries for digital enhancement of images. Accordingly, the court denied the defendant's motion *in limine*.

Nooner v. State of Arkansas, 907 S.W.2d 677 (Ark. 1995)

This is an appeal from a capital murder case where appellant was convicted and sentenced to death by lethal injections. Nooner claims the court erred in admitting "manipulated" photographs into evidence. The State, however, countered that the photographs were only "enhanced" to give more brightness and clarity for better definition and not "altered."

Relevant Case Summary

The victim was shot seven times while at a laundromat. There were three surveillance cameras in the Laundromat, which recorded onto one VHS tape. These cameras were in operation at the time of the shooting, although the actual murder was not captured on tape. The police made still photographs from certain frames from the videotape that showed the victim and the appellant. The tape was enhanced to obtain the clearest still photographs possible. The victim's face was blurred out in several of the enhanced still photos at the request of the victim's family.

Nooner was identified on the surveillance videotape by two witnesses, based on clothing and appearance. Other testimony and ballistics evidence tied Nooner to the murder weapon. The court noted that: "Reliability must be the watchword in determining the admissibility of enhanced videotape and photographs, where by computer or otherwise." *(Id.*, 322 Ark. At 104, 907 S.W.2d at 686)

The court emphasized that there was nothing introduced to indicate that the still photographs introduced into evidence were altered to "include a face, features, or physique of someone not present in the original videotape." In fact the court and jury did see a slowed version of the original tape and could compare it to the enhanced still photos, giving them the opportunity to determine any distortion or irregularity. The appellate court also reviewed the pretrial hearing regarding the photographs and whether they should be suppressed. After reviewing all these facts, the appellate court concluded that there was no abuse of discretion on the issue of allowing the digitally enhanced still photos into evidence.

Additional Brief Overviews of Cases Admitting Digitally Enhanced Photographs from Surveillance Tapes

English v. State

In *English v. State*, 205 Ga. App. 599, 422 S.E.2d 924 (1992), the Georgia Court of Appeals came to the same conclusion and affirmed the admissibility of enhanced photographs taken from surveillance tape. The jury was shown both the original tape and enhanced photo for comparisons. "The …technician who produced the computer-enhanced photographic copy testified to that process and that … the photographic copy, as enhanced, was a 'fair and accurate' representation of what appeared in the … videotape… the computer-enhanced photographic copy would be no less admissible… than the enlargement of a photograph would be." (*Id.*, 205 Ga. App. At 599, 422 S.E.2d at 924. *Accord Rauls v. State*, 209 Ga. App. 101, 102, 432 S.E.2d 677, 678 (1993))

Dolan v. State of Florida

In *Dolan v. State of Florida*, 743 So.2d 544 (Fla. App. 1999), the court admitted a computer-enhanced image from surveillance tapes. The Dolan court stressed the importance of maintaining the original (unenhanced) pristine form for comparison. The Court noted that the proper chain of custody was established through establishing proper operation of the videotape recorder, therefore the issue was "whether the computer stills are verified as reliable representations of images recorded on the original videotape." The court concluded:

> *Once the tape is authenticated and the forensic analyst explains the computer enhancement process and establishes that the images were not altered or edited, then the computer enhancements become admissible as a fair and accurate replicate of what is on the tape, provided the original tape is in evidence for comparison. The weight accorded to those computer-enhanced prints is determined by the jury. A jury possesses sufficient common sense to compare the images.* *Id., 743 So.2d at 546)*

One important factor in many of these cases is that the testifying technician or expert was the person engaged in the enhancement process and had specific knowledge of the detail of the process. Also the testifying expert was capable of explaining in clear language about the process. These cases show the direction that states are headed and what they expect for getting digital evidence into court. There are many tactical pieces of information that experts can gain by reading these types of cases. Learn from them. Pay particular attention to cases that have been decided in your own state. Case law on issues of electronic evidence is continuing to grow and precedent is changing daily. Courts will likely change their requirements for admissibility of electronic evidence as new issues and case law evolves. It would be instructive to keep a keen eye on new case law as it relates to this ever-changing landscape of admissibility of electronic evidence. It is likely that authentication requirements for digitally enhanced photographs will become more stringent, requiring adequate testimony by knowledgeable technicians of the process, tools, methodologies, and related systems that produce the digital evidence. There are adequate and meaningful resources available throughout the forensic community to help maintain the needed processes, methodologies, technical skills, and legal issues. A list of several of these resources can be found at the end of this chapter. Additionally, keeping up with new case law will be as valuable to the expert as any other technical skill in his toolbox. Since the laws and forensic environment are constantly changing, "due diligence" must be the watchword for the future for every forensic expert.

REFERENCES AND RESOURCES

"The Sedona Conference® Commentary on ESI Evidence & Admissibility", The Sedona Conference® Working Group Series, March 2008, free download at http://www.thesedonaconference.org/ditForm?did=ESI_Commentary_0308.pdf – This commentary discusses the existing rules and case law, potential future ESI issues, and a practical guide on using ESI in legal proceedings.

National Institute of Justice. Forensic Examination of Digital Evidence: A Guide for Law Enforcement. April 2004. http://www.ncjrs.gov/pdffiles1/nij/199408.pdf

Electronic Crime Scene Investigation: A Guide for First Responders. July 2001. http://www.ncjrs.gov/pdffiles1/nij/187736.pdf

Digital Evidence in the Courtroom: A Guide for Law Enforcement and Prosecutors. National Institute of Justice. January 2007. http://www.ncjrs.gov/pdffiles1/nij/211314.pdf

Scientific Working Group on Digital Evidence (SWGDE). http://ncfs.ucf.edu/swgde/index.html

Best Practices for Computer Forensics. SWGDE. http://68.156.151.124/documents/swgde2006/Best_Practices_for_Computer_Forensics%20July06.pdf

Scientific Working Group on Imaging Technology (SWGIT). http://www.theiai.org/guidelines/swgit/index.php

SWGIT: Best Practices for Maintaining the Integrity of Digital Images and Digital Video. Draft April 2007. http://www.theiai.org/guidelines/ swgit/guidelines/section_13_v1-0.pdf

National Center for Forensics Science. http://www.ncfs.org

National Center for Forensic Science. Digital Evidence in the Courtroom: A Guide for Preparing Digital Evidence for Courtroom Presentation. http://www.ncfs.org/DE_courtroomdraft.pdf

Electronic Crime Scene Investigation: A Guide for First Responders. http://www.ncjrs.gov/pdffiles1/nij/187736.pdf

US Secret Service. Best Practices for Seizing Electronic Evidence. Available at http://www.secretservice.gov/electronic_evidence.shtml

US Secret Service. Best Practices for Seizing Electronic Evidence V3 a Pocket Guide for First Responders. Available at http://www.forwardedge2.com/pdf/bestPractices.pdf

Forensics Science Communications. Recommendations and Guidelines for the Use of Digital Image Processing in the Criminal Justice System. Scientific Working Group on Imaging Technologies (SWGIT), Version 1.2. June 2002. http://www.fbi.gov/hq/lab/fsc/backissu/jan2003/swgitdigital.htm

Guidelines for Field Applications of Imaging Technologies in the Criminal Justice System. Scientific Working Group on Imaging Technologies (SWGIT), Version 2.3. December 6, 2001. http://www.fbi.gov/hq/lab/fsc/backissu/april2002/swgitfield1.htm.

Institute for Forensic Imaging. http://www.ifi-indy.org/

International Association for Identification (IAI). http://www.theiai.org/

IAI. Guidelines for Working Groups. http://www.theiai.org/guidelines/

SWGDE and SWGIT multimedia glossary. http://ncfs.org/swgde/documents/swgde2006/SWGDE_SWGIT%20Glossary%20V2.0.pdf

SWGDE/SWGIT Guidelines for developing SOP. http://ncfs.org/swgde/documents/swgde2006/SWGDE%20-%20SWGIT%20Recommended%20Guidelines%20for%20Developing%20Standard%20Operating%20Procedures.pdf

SWGDE list of documents. http://ncfs.org/swgde/documents.html

National Forensic Science Technology Center. http://www.nfstc.org/index.htm

US Department of Justice Search and Seizure Manual: Searching and Seizing Computers and Obtaining Electronic Evidence in Criminal Investigations. http://www.justice.gov/criminal/cybercrime/s&smanual2002.pdf.

Susan Flamm, JD, and Samuel H. Solomon. Admissibility of Digital Exhibits in Litigation. Available at http://macproppid.com/documents/Admissibility.pdf.

Rebecca Levy-Sachs and Melissa Sullivan. Using Digital Photographs in the Courtroom-Considerations for Admissibility. Available at http://www.securitymanagement.com/archive/library/feature_August2004.pdf.

Paul R. Dickinson, Jr., and Patricia L. Waddy. How to Get Evidence and Expert Testimony Admitted Into Court, December 14, 2006. Available at http://lewis-roberts.com/downloads/evidence%20&%20Experts.pdf.

Christina Shaw, J.D. Admissibility of Digital Photographic Evidence: Should it be Any Different Than Traditional Photography? Update, Volume 15, Number 10, 2002. Available at http://www.ndaa-apri.org/publications/newsletters/update_volume_15_number_10_2002.html.

SWGIT and SWGIT/SWGDE documents can be found at http://www.theiai.org/swgit/index.html.

International Association of Computer Investigative Science. Forensic Procedure. http://www.iacis.com/iacisv2/pages/forensicprocedures.php

Appendix

Federal Rules of Evidence 2006

Available at http://www.lexisnexis.com/lawschool/learning/reference/pdf/2006/LA11910-0_FRE.pdf

RULE 104. PRELIMINARY QUESTIONS

(a) Questions of admissibility generally. Preliminary questions concerning the qualification of a person to be a witness, the existence of a privilege, or the admissibility of evidence shall be determined by the court, subject to the provisions of subdivision (b). In making its determination it is not bound by the rules of evidence except those with respect to privileges.

(b) Relevancy conditioned on fact. When the relevancy of evidence depends upon the fulfillment of a condition of fact, the court shall admit it upon, or subject to, the introduction of evidence sufficient to support a finding of the fulfillment of the condition.

(c) Hearing of jury. Hearings on the admissibility of confessions shall in all cases be conducted out of the hearing of the jury. Hearings on other preliminary matters shall be so conducted when the interests of justice require, or when an accused is a witness and so requests.

(d) Testimony by accused. The accused does not, by testifying upon a preliminary matter, become subject to cross-examination as to other issues in the case.

(e) Weight and credibility. This rule does not limit the right of a party to introduce before the jury evidence relevant to weight or credibility.

RULE 401. DEFINITION OF "RELEVANT EVIDENCE"

"Relevant evidence" means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.

RULE 402. RELEVANT EVIDENCE GENERALLY ADMISSIBLE; IRRELEVANT EVIDENCE INADMISSIBLE

All relevant evidence is admissible, except as otherwise provided by the Constitution of the United States, by Act of Congress, by these rules, or by other rules prescribed by the Supreme Court pursuant to statutory authority. Evidence which is not relevant is not admissible.

RULE 403. EXCLUSION OF RELEVANT EVIDENCE ON GROUNDS OF PREJUDICE, CONFUSION, OR WASTE OF TIME

Although relevant, evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, or by considerations of undue delay, waste of time, or needless presentation of cumulative evidence.

RULE 701. OPINION TESTIMONY BY LAY WITNESSES

If the witness is not testifying as an expert, the witness' testimony in the form of opinions or inferences is limited to those opinions or inferences which are (a) rationally based on the perception of the witness, and (b) helpful to a clear understanding of the witness' testimony or the determination of a fact in issue, and (c) not based on scientific, technical, or other specialized knowledge within the scope of Rule 702.

RULE 702. TESTIMONY BY EXPERTS

If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

RULE 703. BASES OF OPINION TESTIMONY BY EXPERTS

The facts or data in the particular case upon which an expert bases an opinion or inference may be those perceived by or made known to the expert at or before the hearing. If of a type reasonably relied upon by experts in the particular field in forming opinions or inferences upon the subject, the facts or data need not be admissible in evidence in order for the opinion or inference to be admitted. Facts or data that are otherwise inadmissible shall not be disclosed to the jury by the proponent of the opinion or inference unless the court determines that their probative value in assisting the jury to evaluate the expert's opinion substantially outweighs their prejudicial effect.

RULE 704. OPINION ON ULTIMATE ISSUE

(a) Except as provided in subdivision (b), testimony in the form of an opinion or inference otherwise admissible is not objectionable because it embraces an ultimate issue to be decided by the trier of fact.

(b) No expert witness testifying with respect to the mental state or condition of a defendant in a criminal case may state an opinion or inference as to whether the defendant did or did not have the mental state or condition constituting an element of the crime charged or of a defense thereto. Such ultimate issues are matters for the trier of fact alone.

RULE 705. DISCLOSURE OF FACTS OR DATA UNDERLYING EXPERT OPINION

The expert may testify in terms of opinion or inference and give reasons therefore without first testifying to the underlying facts or data, unless the court requires otherwise. The expert may in any event be required to disclose the underlying facts or data on cross-examination.

RULE 706. COURT APPOINTED EXPERTS

(a) Appointment. The court may on its own motion or on the motion of any party enter an order to show cause why expert witnesses should not be appointed, and may request the parties to submit nominations. The court may appoint any expert witnesses agreed upon by the parties, and may appoint expert witnesses of its own selection. An expert witness shall not be appointed by the court unless the witness consents to act. A witness so appointed shall be informed of the witness' duties by the court in writing, a copy of which shall be filed with the clerk, or at a conference in which the parties shall have opportunity to participate. A witness so appointed shall advise the parties of the witness' findings, if any; the witness' deposition may be taken by any party; and the witness may be called to testify by the court or any party. The witness shall be subject to cross-examination by each party, including a party calling the witness.

(b) Compensation. Expert witnesses so appointed are entitled to reasonable compensation in whatever sum the court may allow. The compensation thus fixed is payable from funds which may be provided by law in criminal cases and civil actions and proceedings involving just compensation under the fifth amendment. In other civil actions and proceedings the compensation shall be paid by the parties in such proportion and at such time as the court directs, and thereafter charged in like manner as other costs.

(c) Disclosure of appointment. In the exercise of its discretion, the court may authorize disclosure to the jury of the fact that the court appointed the expert witness.

(d) Parties' experts of own selection. Nothing in this rule limits the parties in calling expert witnesses of their own selection.

RULE 901. REQUIREMENT OF AUTHENTICATION OR IDENTIFICATION

(a) General provision. The requirement of authentication or identification as a condition precedent to admissibility is satisfied by evidence sufficient to support a finding that the matter in question is what its proponent claims.

(b) Illustrations. By way of illustration only, and not by way of limitation, the following are examples of authentication or identification conforming with the requirements of this rule:

(1) Testimony of witness with knowledge. Testimony that a matter is what it is claimed to be.

(2) Nonexpert opinion on handwriting. Nonexpert opinion as to the genuineness of handwriting, based upon familiarity not acquired for purposes of the litigation.

(3) Comparison by trier or expert witness. Comparison by the trier of fact or by expert witnesses with specimens which have been authenticated.

(4) Distinctive characteristics and the like. Appearance, contents, substance, internal patterns, or other distinctive characteristics, taken in conjunction with circumstances.

(5) Voice identification. Identification of a voice, whether heard firsthand or through mechanical or electronic transmission or recording, by opinion based upon hearing the voice at any time under circumstances connecting it with the alleged speaker.

(6) Telephone conversations. Telephone conversations, by evidence that a call was made to the number assigned at the time by the telephone company to a particular person or business, if (A) in the case of a person, circumstances, including self-identification, show the person answering to be the one called, or (B) in the case of a business, the call was made to a place of business and the conversation related to business reasonably transacted over the telephone.

(7) Public records or reports. Evidence that a writing authorized by law to be recorded or filed and in fact recorded or filed in a public office, or a purported public record, report, statement, or data compilation, in any form, is from the public office where items of this nature are kept.

(8) Ancient documents or data compilation. Evidence that a document or data compilation, in any form, (A) is in such condition as to create no suspicion concerning its authenticity (B) was in a place where it, if authentic, would likely be, and (C) has been in existence 20 years or more at the time it is offered.

(9) Process or system. Evidence describing a process or system used to produce a result and showing that the process or system produces an accurate result.

(10) Methods provided by statute or rule. Any method of authentication or identification provided by Act of Congress or by other rules prescribed by the Supreme Court pursuant to statutory authority.

RULE 902. SELF-AUTHENTICATION

Extrinsic evidence of authenticity as a condition precedent to admissibility is not required with respect to the following:

(1) Domestic public documents under seal. A document bearing a seal purporting to be that of the United States, or of any State, district, Commonwealth, territory, or insular possession thereof, or the Panama Canal Zone, or the Trust Territory of the Pacific Islands, or of a political subdivision, department, officer, or agency thereof, and a signature purporting to be an attestation or execution.

(2) Domestic public documents not under seal. A document purporting to bear the signature in the official capacity of an officer or employee of any entity included in paragraph (1) hereof, having no seal, if a

public officer having a seal and having official duties in the district or political subdivision of the officer or employee certifies under seal that the signer has the official capacity and that the signature is genuine.

(3) Foreign public documents. A document purporting to be executed or attested in an official capacity by a person authorized by the laws of a foreign country to make the execution or attestation, and accompanied by a final certification as to the genuineness of the signature and official position (A) of the executing or attesting person, or (B) of any foreign official whose certificate of genuineness of signature and official position relates to the execution or attestation or is in a chain of certificates of genuineness of signature and official position relating to the execution or attestation. A final certification may be made by a secretary of an embassy or legation, consul general, consul, vice consul, or consular agent of the United States, or a diplomatic or consular official of the foreign country assigned or accredited to the United States. If reasonable opportunity has been given to all parties to investigate the authenticity and accuracy of official documents, the court may, for good cause shown, order that they be treated as presumptively authentic without final certification or permit them to be evidenced by an attested summary with or without final certification.

(4) Certified copies of public records. A copy of an official record or report or entry therein, or of a document authorized by law to be recorded or filed and actually recorded or filed in a public office, including data compilations in any form, certified as correct by the custodian or other person authorized to make the certification, by certificate complying with paragraph (1), (2), or (3) of this rule or complying with any Act of Congress or rule prescribed by the Supreme Court pursuant to statutory authority.

(5) Official publications. Books, pamphlets, or other publications purporting to be issued by public authority.

(6) Newspapers and periodicals. Printed materials purporting to be newspapers or periodicals.

(7) Trade inscriptions and the like. Inscriptions, signs, tags, or labels purporting to have been affixed in the course of business and indicating ownership, control, or origin.

(8) Acknowledged documents. Documents accompanied by a certificate of acknowledgment executed in the manner provided by law by a notary public or other officer authorized by law to take acknowledgments.

(9) Commercial paper and related documents. Commercial paper, signatures thereon, and documents relating thereto to the extent provided by general commercial law.

(10) Presumptions under Acts of Congress. Any signature, document, or other matter declared by Act of Congress to be presumptively or prima facie genuine or authentic.

(11) Certified domestic records of regularly conducted activity. The original or a duplicate of a domestic record of regularly conducted activity that would be admissible under Rule 803(6) if accompanied by a written declaration of its custodian or other qualified person, in a manner complying with any Act of Congress or rule prescribed by the Supreme Court pursuant to statutory authority, certifying that the record:

(A) was made at or near the time of the occurrence of the matters set forth by, or from information transmitted by, a person with knowledge of those matters;

(B) was kept in the course of the regularly conducted activity; and

(C) was made by the regularly conducted activity as a regular practice.

A party intending to offer a record into evidence under this paragraph must provide written notice of that intention to all adverse parties, and must make the record and declaration available for inspection sufficiently in advance of their offer into evidence to provide an adverse party with a fair opportunity to challenge them.

(12) Certified foreign records of regularly conducted activity. In a civil case, the original or a duplicate of a foreign record of regularly conducted activity that would be admissible under Rule 803(6) if accompanied by a written declaration by its custodian or other qualified person certifying that the record:

(A) was made at or near the time of the occurrence of the matters set forth by, or from information transmitted by, a person with knowledge of those matters;

(B) was kept in the course of the regularly conducted activity; and

(C) was made by the regularly conducted activity as a regular practice.

The declaration must be signed in a manner that, if falsely made, would subject the maker to criminal penalty under the laws of the country where the declaration is signed. A party intending to offer a record into evidence under this paragraph must provide written notice of that intention to all adverse parties, and must make the record and declaration available for inspection sufficiently in advance of their offer into evidence to provide an adverse party with a fair opportunity to challenge them.

RULE 903. SUBSCRIBING WITNESS' TESTIMONY UNNECESSARY

The testimony of a subscribing witness is not necessary to authenticate a writing unless required by the laws of the jurisdiction whose laws govern the validity of the writing.

RULE 1001. DEFINITIONS

For purposes of this article the following definitions are applicable:

(1) Writings and recordings. "Writings" and "recordings" consist of letters, words, or numbers, or their equivalent, set down by handwriting, typewriting, printing, photostating, photographing, magnetic impulse, mechanical or electronic recording, or other form of data compilation.

(2) Photographs. "Photographs" include still photographs, X-ray films, video tapes, and motion pictures.

(3) Original. An "original" of a writing or recording is the writing or recording itself or any counterpart intended to have the same effect by a person executing or issuing it. An "original" of a photograph includes the negative or any print therefrom. If data are stored in a computer or similar device, any printout or other output readable by sight, shown to reflect the data accurately, is an "original".

(4) Duplicate. A "duplicate" is a counterpart produced by the same impression as the original, or from the same matrix, or by means of photography, including enlargements and miniatures, or by mechanical or electronic re-recording, or by chemical reproduction, or by other equivalent techniques which accurately reproduces the original.

RULE 1002. REQUIREMENT OF ORIGINAL

To prove the content of a writing, recording, or photograph, the original writing, recording, or photograph is required, except as otherwise provided in these rules or by Act of Congress.

Notes of Advisory Committee on Rules for Rule 1002

The assumption should not be made that the rule will come into operation on every occasion when use is made of a photograph in evidence. On the contrary, the rule will seldom apply to ordinary photographs. In most instances a party wishes to introduce the item and the question raised is the propriety of receiving it in evidence. Cases in which an offer is made of the testimony of a witness as to what he saw in a photograph or motion picture, without producing the same, are most unusual. The usual course is for a witness on the stand to identify the photograph or motion picture as a correct representation of events which he saw or of a scene with which he is familiar. In fact he adopts the picture as his testimony, or, in common parlance, uses the picture to illustrate his testimony. Under these circumstances, no effort is made to prove the contents of the picture, and the rule is inapplicable. Paradis, The Celluloid Witness, 37 U.Colo.L. Rev. 235, 249–251 (1965).

On occasion, however, situations arise in which contents are sought to be proved. Copyright, defamation, and invasion of privacy by photograph or motion picture falls in this category. Similarly, as for situations in which the picture is offered as having independent probative value, e.g. automatic photograph of bank robber. See *People v. Doggett*, 83 Cal.App.2d 405, 188 P.2d 792 (1948) photograph of defendants engaged in indecent act; Mouser and Philbin, Photographic Evidence – Is There a Recognized Basis for Admissibility? 8 Hastings L.J. 310 (1957). The most commonly encountered of this latter group is of course, the X-ray, with substantial authority calling for production of the original. *Daniels v. Iowa City,* 191 Iowa 811, 183 N.W. 415 (1921); *Cellamare v. Third Acc. Transit Corp.,* 273 App.Div. 260, 77 N.Y.S.2d 91 (1948); *Patrick & Tilman v. Matkin,* 154 Okl. 232, 7 P.2d 414 (1932); *Mendoza v. Rivera,* 78 P.R.R. 569 (1955).

RULE 1003. ADMISSIBILITY OF DUPLICATES

A duplicate is admissible to the same extent as an original unless (1) a genuine question is raised as to the authenticity of the original or (2) in the circumstances it would be unfair to admit the duplicate in lieu of the original.

RULE 1007. TESTIMONY OR WRITTEN ADMISSION OF PARTY

Contents of writings, recordings, or photographs may be proved by the testimony or deposition of the party against whom offered or by that party's written admission, without accounting for the non-production of the original.

REFERENCES AND RESOURCES

National Institute of Justice. Forensic Examination of Digital Evidence: A Guide for Law Enforcement. April 2004. http://www.ncjrs.gov/pdffiles1/nij/199408.pdf

This guide is intended for use by members of the law enforcement community who are responsible for the examination of digital evidence. The guide, published as an NIJ Special Report, is the second in a series of guides on investigating electronic crime (see also *Electronic Crime Scene Investigation: A Guide for First Responders*). It deals with common situations encountered during the processing and handling of digital evidence and can be used to help agencies develop their own policies and procedures.

Electronic Crime Scene Investigation. A Guide for First Responders. July 2001. http://www.ncjrs.gov/pdffiles1/nij/187736.pdf

Digital Evidence in the Courtroom: A Guide for Law Enforcement and Prosecutors. National Institute of Justice. January 2007. http://www.ncjrs.gov/pdffiles1/nij/211314.pdf

Summary: Now essential to modern life, computers have also become increasingly important to criminals, who steal information, commit fraud, and stalk victims online. Even if a crime was not committed online, law enforcement may discover critical evidence from an offenders' digital media. For this evidence to be admissible, however, police must demonstrate proper collection and handling. In the courtroom, prosecutors must overcome the twin barriers of skepticism and lack of technical understanding. To help navigate this complex process, NIJ's technical working group of national experts prepared this special report. Chapters 1 and 2 inform crime scene examiners and other handlers about legal requirements for the handling of digital evidence. Chapters 3 and 4 provide guidelines for successful prosecution. The last chapter is a working application—using digital evidence to convict in a child pornography case. Appendixes provide useful resources and forms.

Scientific Working Group on Digital Evidence (SWGDE). http://ncfs.ucf.edu/swgde/index.html

Best Practices for Computer Forensics. SWGDE. http://ncfs.org/swgde/documents/swgde2006/Best_Practices_for_Computer_Forensics%20July06.pdf

Scientific Working Group on Imaging Technology (SWGIT). http://www.theiai.org/guidelines/swgit/index.php

SWGIT. Best Practices for Maintaining the Integrity of Digital Images and Digital Video. Draft, April 2007. http://www.theiai.org/guidelines/swgit/guidelines/section_13_v1-0.pdf

National Center for Forensic Science. http://www.ncfs.org

National Center for Forensic Science. Digital Evidence in the Courtroom: A Guide for Preparing Digital Evidence for Courtroom Presentation. http://www.ncfs.org/DE_courtroomdraft.pdf

Electronic Crime Scene Investigation: A Guide for First Responders. http://www.ncjrs.gov/pdffiles1/nij/187736.pdf

Forensics Science Communications. Recommendations and Guidelines for the Use of Digital Image Processing in the Criminal Justice System. Scientific Working Group on Imaging Technologies (SWGIT), Version 1.2. June 2002. http://www.fbi.gov/hq/lab/fsc/backissu/jan2003/swgitdigital.htm

Guidelines for Field Applications of Imaging Technologies in the Criminal Justice System. Scientific Working Group on Imaging Technologies (SWGIT), Version 2.3. December 6, 2001.
http://www.fbi.gov/hq/lab/fsc/backissu/april2002/swgitfield1.htm

Institute for Forensic Imaging. http://www.ifi-indy.org/

International Association for Identification (IAI). http://www.theiai.org/

IAI. Guidelines for Working Groups. http://www.theiai.org/guidelines/

SWGDE and SWGIT multimedia glossary. http://ncfs.org/swgde/documents/swgde2006/SWGDE_SWGIT%20Glossary%20V2.0.pdf

SWGDE/SWGIT Guidelines for developing SOP.http://ncfs.org/swgde/documents/swgde2006/SWGDE%20-%20SWGIT%20Recommended%20Guidelines%20for%20Developing%20Standard%20Operating%20Procedures.pdf

SWGDE list of documents. http://ncfs.org/swgde/documents.html

National Forensic Science Technology Center. http://www.nfstc.org/index.htm

US Department of Justice Search and Seizure Manual. Searching and Seizing Computers and Obtaining Electronic Evidence in Criminal Investigations.
http://www.justice.gov/criminal/cybercrime/s&smanual2002.pdf

Susan Flamm, JD, and Samuel H. Solomon. Admissibility of Digital Exhibits in Litigation. Available at http://macproppid.com/documents/Admissibility.pdf.

Rebecca Levy-Sachs and Melissa Sullivan. Using Digital Photographs in the Courtroom—Considerations for Admissibility. Available at http://www.securitymanagement.com/archive/library/feature_August2004.pdf.

Paul R. Dickinson, Jr., and Patricia L. Waddy. How to Get Evidence and Expert Testimony Admitted Into Court. December 14, 2006. Available at http://lewis-roberts.com/downloads/evidence%20&%20Experts.pdf.

Christina Shaw, J.D. Admissibility of Digital Photographic Evidence: Should it be Any Different Than Traditional Photography? Update, Volume 15. Number 10, 2002. Available at http://www.ndaa-apri.org/publications/newsletters/update_volume_15_number_10_2002.html.

REFERENCE LIST

SWGIT and SWGIT/SWGDE documents can be found at http://www.theiai.org/swgit/index.html.

Section	Title
Section 1	Overview of SWGIT and the Use of Imaging Technology in the Criminal Justice System
Section 2	Considerations for Managers Migrating to Digital Imaging Technology
Section 3	Guidelines for Field Applications of Imaging Technologies in the Criminal Justice System
Section 4	Recommendations and Guidelines for Using Closed-Circuit Television Security Systems in Commercial Institutions
Section 5	Recommendations and Guidelines for the Use of Digital Image Processing in the Criminal Justice System
Section 6	Guidelines and Recommendations for Training in Imaging Technologies in the Criminal Justice System
Section 7	Recommendations and Guidelines for the Use of Forensic Video Processing in the Criminal Justice System
Section 8	General Guidelines for Capturing Latent Impressions Using a Digital Camera
Section 9	General Guidelines for Photographing Tire Impressions
Section 10	General Guidelines for Photographing Footwear Impressions
Section 11	Best Practices for Documenting Image Enhancement
Section 12	Best Practices for Forensic Image Analysis
Section 13	Best Practices for Maintaining the Integrity of Digital Images and Digital Video
Section 14	Best Practices for Image Authentication
Section 15	Best Practices for Archiving Digital and Multimedia Evidence in the Criminal Justice System

SWGIT/ SWGDE	Proficiency Test Program Guidelines
SWGIT/ SWGDE	Guidelines and Recommendations for Training in Digital and Multimedia Evidence
SWGIT/ SWGDE	Recommended Guidelines for Developing Standard Operating Procedures
SWGIT/ SWGDE	Glossary of Terms

International Association of Computer Investigative Science—Forensic Procedure http://www.iacis.com/iacisv2/pages/forensicprocedures.php

Index