CHAPTER 4

Multi-Modal Classifier-Fusion for the Recognition of Emotions

Martin Schels, Michael Glodek, Sascha Meudt, Stefan Scherer, Miriam Schmidt, Georg Layher, Stephan Tschechne, Tobias Brosch, David Hrabal, Steffen Walter, Harald C. Traue, Günther Palm, Heiko Neumann and Friedhelm Schwenker

1. Introduction and Motivation

Research activities in the field of human-computer interaction increasingly addressed the aspect of integrating features that characterize different types of emotional *intelligence*. Human emotions are expressed through different modalities such as speech, facial expressions, hand or body gestures, and therefore the classification of human emotions should be considered as a multi-modal pattern recognition problem. In recent time, a multitude of approaches have been proposed to enhance the training and recognition of multiple classifier systems (MCSs) utilizing multiple modalities to classify human emotional states. The work summarizes the progress of investigating such systems and presents aspects of the problem namely fusion architectures and training of statistical classifiers based on only marginal informative features. Furthermore, it describes how the effects of missing values, e.g. due to missing sensor data or classifiers exploiting reject options in order to circumvent false classifications, can be mitigated. Another aspect is the usage of partially supervised learning, either to support annotation or to improve classifiers. Parts of these aspects are then exemplified using two recent examples of emotion recognition, showing a successful realization of MCSs in emotion recognition.

Research in affective computing has made many achievements in the last years. Emotions begin to play an increasingly important role in the field of human-computer interaction, allowing the user to interact with the system more efficiently (Picard, 2003) and in a more natural way (Sebe et al., 2007). Such a system must be able to recognize the users' emotional state, which can be done by analyzing the facial expression (Ekman and Friesen, 1978), taking the body posture and the gestures into account (Scherer et al., 2012) and by investigating the paralinguistic information hidden in the speech (Schuller et al., 2003; Oudeyer, 2003). Furthermore, biophysiological channels can provide valuable information to conclude to the affective state (Cannon, 1927; Schachter, 1964).

However, the emotions investigated so far were in general acted and the larger part of research was focused on a single modality, albeit the problem of emotion recognition is inherently multi-modal. Obviously, the entire emotional state of an individual is expressed and can be observed in different modalities, e.g. through facial expressions, speech, prosody, body movement, hand gestures as well as more internal signals such as heart rate, skin conductance, respiration, electroencephalography (EEG) or electromyogram (EMG). Recent developments aim at transferring the insights obtained from single modalities and acted emotions to more natural settings using multiple modalities (Caridakis et al., 2007; Scherer et al., 2012; Zeng et al., 2009; Chen and Huang, 2000). The uncontrolled recording of non-acted data and the manifold of modalities make emotion recognition a challenging task: subjects are less restricted in their behavior, emotions occur more rarely and the emotional ground truth is difficult to determine, because human observers also tend to disagree about emotions.

In this chapter, MCSs for the classification of multi-modal features are investigated, the numerical evaluation of the proposed emotion recognition systems is carried out on the data sets of the 1st AVEC challenge (Schuller et al., 2011) and a data set recorded in a Wizard-of-Oz scenario (Walter et al., 2011). Combining multi-modal classifiers is a promising approach to improve the overall classifier performance (Schels and Schwenker, 2010). Such a team of classifiers should be accurate and diverse (Kuncheva, 2004). While the requirement to the classifiers to be as accurate as possible is obvious, diversity roughly means that classifiers should not agree on misclassified data. In our studies, various modalities and feature views have been utilized on the data to achieve such a set of diverse and accurate classifiers.

The rest of this chapter is organized as follows: In Section 2, we will present the latest approaches to improve the recognition of emotion. Section 3 describes real-world data collections for affective computing. Furthermore, adequate features are described together with a numerical evaluation. Finally, Section 4 concludes.

2. Multi-modal Classification Architectures and Information Fusion for Emotion Recognition

2.1 Learning from multiple sources

For many benchmark data collections in the field of machine learning, it is sufficient to process one type of feature that is extracted from a single representation of the data (e.g. visual digit recognition). However, often in many real-world applications, different independent sensors are available (e.g. microphone and camera) and it is necessary to combine these channels to obtain a good recognition performance and to achieve a robust architecture against sensor failure.

To create a classifier system, which is able to handle different sources of information, three widely used approaches have been proposed and evaluated in the literature, namely *early fusion*, *mid-level fusion* and *late fusion* (Dietrich et al., 2003). Using early fusion, the information is combined on the earliest level by concatenating the individual features to a higher dimensional vector, as depicted on the left-hand side of Figure 1. The converse strategy is to combine the independent streams as late as possible, which is called late fusion or multiple classifier system (MCS), see the right-hand side of Figure 1. The third approach, which recently gains more attention, is known as mid-level fusion (Scherer et al., 2012; Eyben et al., 2012; Glodek et al., 2012; Dietrich et al., 2003) and combines the channels in an intermediate abstraction level, as for example conducted in a combined hidden layer of an artificial neural network. The corresponding classifier architecture is shown in the middle of Figure 1.

The selection of an optimal architecture is strongly related to the respective problem. An important clue for choosing the appropriate architecture could be drawn by judging the dependency and

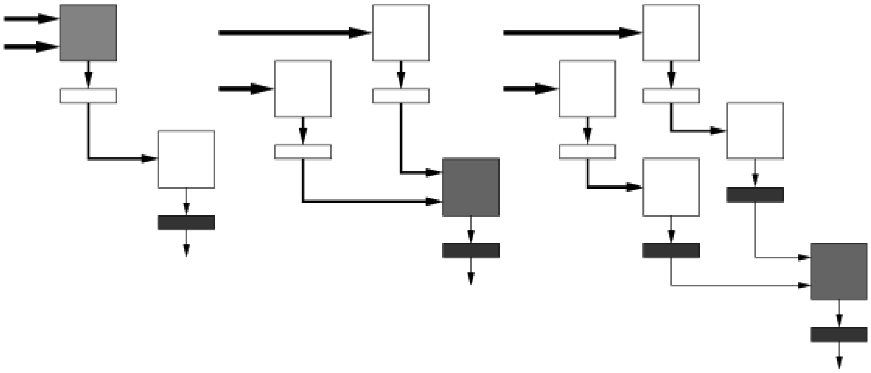

Figure 1. Schematic depiction of different classifier architectures: early fusion, mid-level fusion and late fusion (left to right).

information-gain of the features, and the complexity of the classifier function. Concatenating features of different sources is advantageous because the classification task may become separable. However, extending the dimensionality also implicates to run into the so-called *curse of dimensionality* (Bishop, 2006). Furthermore, in the application of emotion recognition, early fusion is not intuitive as the individual sources are likely to have different sampling rates.

Further, it is often necessary to compensate failing sensors that may occur for example when subjects move away from the camera or when physiological sensors lose contact to the subject's skin. Hence, it is intuitive to combine the individual features as late as possible in an abstract representation.

The mid-level fusion is a good compromise between the two extremes. Figure 2 shows a layered classifier architecture for recognizing long-term user categories. According to the key concept, the patterns are always classified based on the output of the proceeding layer such that the temporal granularity likewise the level of abstractness constantly increases. According to the theory, the architecture is able to recognize classes which are not directly observable (e.g. the affective state) based on the available evidences (Glodek et al., 2011; Scherer et al., 2012).

MCSs are widely used in the machine learning community (Kuncheva, 2004). The performance of an MCS not only depends on the accuracies of the individual classifiers, but also on the diversity of the classifiers, which roughly means that classifiers should not agree on the set of misclassified data. MCSs are highly efficient pattern recognizers that have been studied by various numerical experiments and mathematical analysis, and lead to numerous practical applications

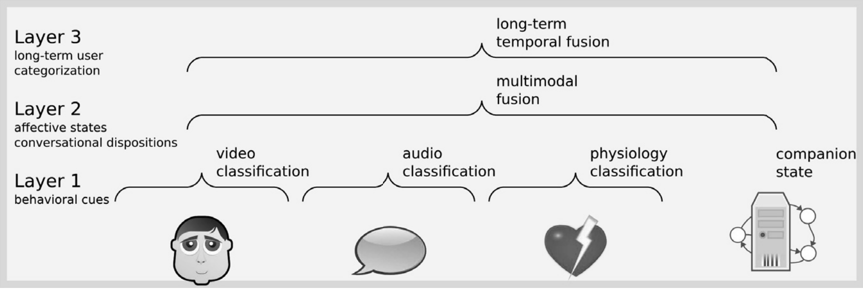

Figure 2. Layered (mid-level) classification architecture to recognize dispositions in human-companion interaction. The level of abstraction increases in each layer to obtain high-level symbolic information.

such as activity recognition (Glodek et al., 2012), EEG analysis (Schels et al., 2011) and classification of bio-acoustic signals (Dietrich et al., 2003) to mention just a few examples. There are different techniques in the literature to attain diverse ensemble classifiers. The individual classifiers can, for example, be trained on different subsets of the training data (Breiman, 1996). Another way is to conduct multiple training runs on the data using different base models or different configurations of a model (model averaging). Furthermore, different subsets of the available feature space (so-called feature views) are often used to construct individual classifiers, which are then treated as independent data streams.

In order to formally reflect that the accuracies of individual classifiers in real-world scenarios and especially in non-acted affective computing are generally low, it is useful to implement mechanisms to increase the robustness and to assess the quality of a decision for a sample. While the robustness can be achieved using the aforementioned ensembles of classifiers, the self-assessment of classifiers can be obtained by defining an appropriate uncertainty measure. Common ways to establish uncertainty measures are to use probabilistic or fuzzy classifiers, or to use the degree of agreement of an ensemble of classifiers, i.e. the more the individual classifiers agree on a specific value or label, the more confident this decision can be seen as. When combining multiple decisions, the uncertainty can be used as a weight in the fusion (Glodek et al., 2012).

Especially in real-world scenarios, it has been proven to be successful to stabilize weak decisions by integrating individual results over time (Glodek et al., 2012). Hereby, the confidence of the classifier can also help to assess weak decisions. This integration could also slow down the sample rates to match the sample rates of the sensory channels.

Human emotions occur in many variations and are often not directly accessible even for human experts when annotating affective corpora. Hence, a severe issue in affective computing is that the labeling procedure is inevitably expensive and time consuming. It would be desirable to incorporate unlabeled data in the overall classification process. This can be done either to improve a statistical learning process or to support a human expert in an interactive labeling process (Meudt et al., 2012). In order to integrate unlabeled data in a supervised machine learning procedure, two different partially supervised learning approaches have been applied, namely *semi-supervised learning* and *active learning*. *Semi-supervised learning* refers to group of methods that attempt to take advantage of unlabeled data for supervised learning (*semi-supervised classification*) or to incorporate prior information such as class labels, pair-wise constraints or cluster membership (*semi-supervised clustering*). *Active learning* or *selective sampling* (Settles, 2009) refers to methods where the learning algorithm has control on the data selection, e.g. it can select the most important/informative examples from a pool of unlabeled examples, then a human expert is asked for the correct data label. The aim is to reduce annotation costs. In our application—the recognition of human emotions in human-computer interaction—we focus more on active learning (Schwenker and Trentin, 2012; Abdel Hady and Schwenker, 2010). An iterative labeling process is displayed in Figure 3, where a machine classifier proposes labels for different areas in a recording for an expert to acknowledge. Based on this, new propositions can be made by the system.

In affective computing, it is not likely that it is necessary to make a decision for every given data sample extracted from a short time analysis. Additionally data samples are delivered relatively often compared to the expected lengths of the observed categories. Hence, it is intuitive to use techniques of sample rejection, i.e. deciding (yes or no) whether a certain confidence level has been achieved or not. Various attempts have been made to introduce confidence-based rejection criteria. Commonly thresholds-based heuristics are used on probabilistic classifier outputs utilizing a distinct uncertainty calculus, for instance doubt and conflict values computed through Dempster's rule of combination in the very well-known Dempster-Shafer theory of evidence (Thiel et al., 2005). Fusion architectures which are making use of reject options not only have to deal with missing signals and different sample rates but also with missing decisions due to rejection.

For these reasons and also as mentioned above, a classification architecture that is designed for a real-world application has to

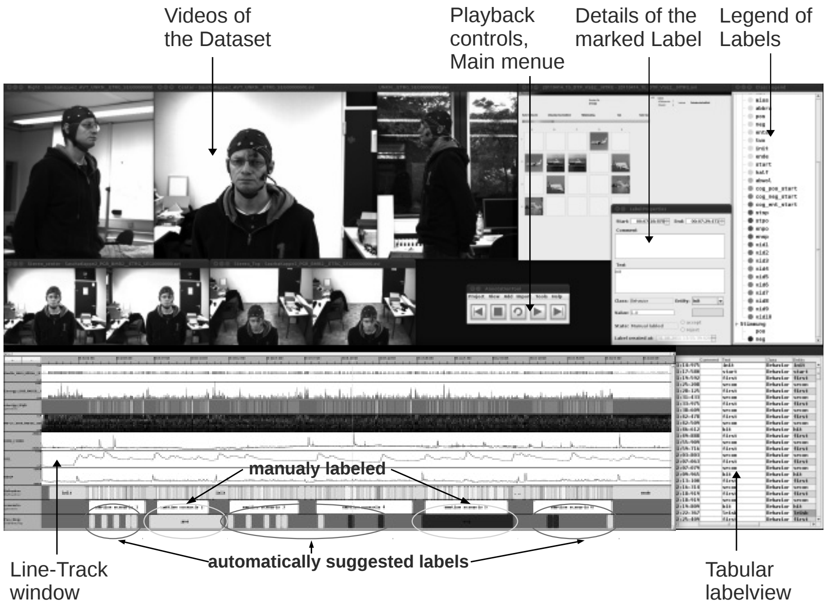
Videos of the Dataset
Playback controls, Main menue
Details of the marked Label
Legend of Labels
manualy labeled
automatically suggested labels
Line-Track window
Tabular labelview

Figure 3. Depiction of the individual channels in an affectively colored human-computer interaction. Furthermore, a semi-automatic annotation process is depicted: The labels encircled in green are set by the human annotator and the system proposes the red labels additionally.

(Color image of this figure appears in the color plate section at the end of the book.)

be robust against missing data. This can be achieved by temporal smoothing of the results as seen in Figure 4. There, the blue short lines represent decisions for the audio channel in word granularity, the orange dots represent decisions for video frames and the short green lines are decisions based on physiological signals such as skin conductance. When a line is drawn in a lighter color, the sample is rejected due to a low respective confidence. But still a decision for every time step is returned by exploiting the hypothesis that the lateral differences over time are low. Further, it is possible to stack multiple layers of classifiers in order to assess more complex categories based on simpler observations. When using, for example, statistical models that can incorporate time series, this architecture can reflect high level concepts that are not directly observable in the data.

Based on this, we propose a classification architecture, as depicted in Figure 5, for the recognition of affective states in human-computer interaction. Here, every individual channel is classified separately with the usage of an uncertainty measure. Based on this, a sample reject mechanism is applied in order to prevent false classifications. The subsequent temporal integration is used not only to further improve

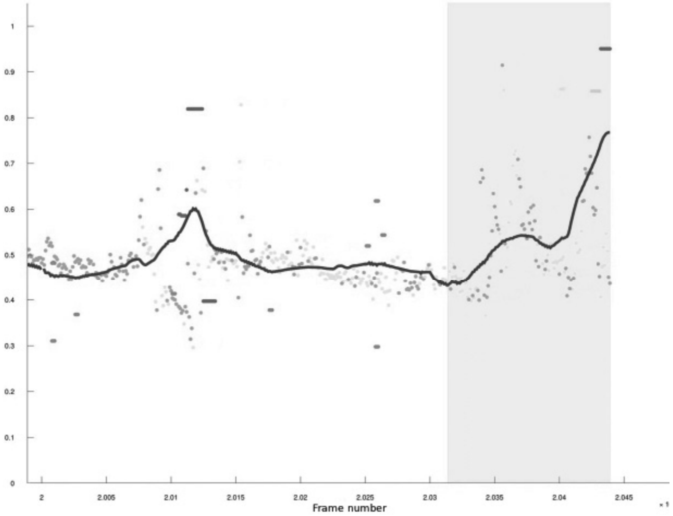

Figure 4. Example for the classification of arousal in human-computer interaction. The strong colors represent decisions that were used in the fusion (solid line), whereas the lighter dots represent rejected samples.

(Color image of this figure appears in the color plate section at the end of the book.)

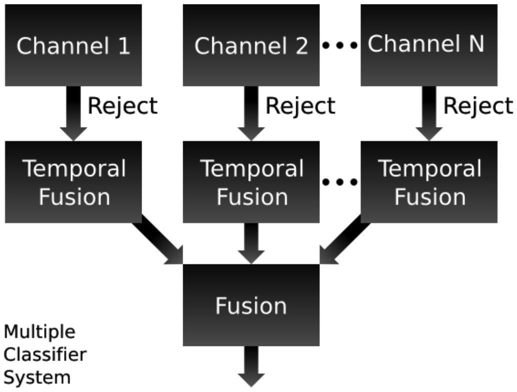

Figure 5. Multiple classifier architecture, which is making use of the reject option. The classification result of each channel has to pass a rejection step, in which decisions with low confidences are filtered out. The outcome is temporally fused and combined.

classification but also to reconstruct missing values. Finally a classifier combination is conducted.

2.2 Base classifiers

In our architectures, linear classifiers, artificial neural networks (e.g. multilayer perceptrons (MLPs)) and support vector machines (SVMs) were used as base classifiers. Linear classifiers are advantageous for noisy data and to avoid over-fitting, because the decisions are based on a rather simple function, namely the linear combination of the features. We obtained the mapping by computing the Moore-Penrose pseudo-inverse function. MLPs are based on a superposition of multiple functions (e.g. linear or sigmoid functions), which are represented by the neurons in the hidden layer (Haykin, 1999). As a result, the complexity of the MLP can be conveniently adjusted by varying the number of hidden neurons.

The SVM is a supervised learning method following the maximum margin paradigm. The classical implementation of the SVM is a typical representative of a kernel method, and therefore the so-called kernel trick can be applied. The kernel trick conducts a mapping to a new feature space that allows classification using non-linear hyper-planes. Within our study, we used the Gaussian radial basis function (RBF) kernel, which transforms the input data into the Hilbert space of infinite dimensions and is calibrated by a width parameter. However, due to noise or incorrect annotations, it is convenient to have a non-rigid hyper-plane, being less sensitive to outliers in the training. Therefore, an extension to the SVM introduces a so-called slack term that tolerates the amount of misclassified data using the control parameter. A probabilistic classification output can be obtained using the method proposed in Platt (1999). Detailed information of these algorithms can be found for instance in Bishop (2006).

Furthermore, Markov models such as the hidden Markov model (HMM) have proven to be a suited method for emotion algorithms (Glodek et al., 2011). The HMM is a stochastic model, applied for temporal/sequential pattern recognition, e.g. speech recognition and recognition of gestures. It is composed of two random processes, a Markov chain with hidden states for the transitions through the states and a second random process modeling the observations. The transition probabilities and also the emission probabilities for the outputs are estimated using the Baum-Welch algorithm. Given the parameters of an HMM and an observed output sequence, the most likely state sequence can be computed (Viterbi algorithm) and the posteriori probability for the observation sequence can be estimated (forward algorithm). This probability can be utilized to classify sequences, by choosing the most likely model (Rabiner, 1989).

3. Experiments

In the following section, several of the aspects presented in the previous sections are evaluated using non-acted emotional data sets.

3.1 Data collection

We will put focus on two different data sets, namely the EmoRec II collected at the Ulm University and the AVEC 2011 dataset.

3.1.1 EmoRec II

A simulation of a natural verbal human-computer interaction was implemented as a Wizard-of-Oz (WOZ) experiment (Kelley, 1984). The WOZ experiment allows the simulation of computer or system properties in a manner such that subjects have the impression that they are having a completely natural verbal interaction with a computer-based memory trainer. The design of the memory trainer followed the principle of the popular game "Concentration". The variation of the system behavior, in response to the subjects, was implemented via natural spoken language, with parts of the subject's reactions taken automatically into account.

In order to induce target emotions during the experiment, we considered the following affective factors that are implemented as natural language dialog:

- Delaying the response of a command
- Non-execution of the command
- Simulate incorrect speech recognition
- Offer of technical assistance
- Lack of technical assistance
- Propose to quit the game ahead of time
- Positive feedback

The procedure of emotion induction is structured in differentiated experimental sequences (ESs) in which the user is passed through VAD octants (valence: positive, negative, neutral; arousal: low, high, neutral; dominance: low control, high control, neutral) by the investigator (compare Figure 6). Audio, video and physiological data (namely electromyography, skin conductance level) were recorded.

Within this study, we focus on the recognition of the emotional octants in ES-4 and ES-6. The database comprises eight subjects with an average age of 63.5 years.

3.1.2 AVEC 2011 Data Collection

The second data collection used in this study has been provided within the Audio/Visual Emotion Challenge (AVEC) 2011 of the ACII 2011 workshop. Overall three sub-challenges were proposed: an audio

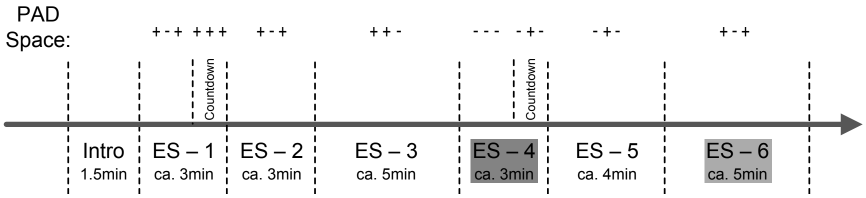

Figure 6. Experimental sequences the subject is guided through for the recordings.

challenge on word-level, a video challenge on frame-level and an audiovisual also on video frame-level.

The data was recorded in a human-computer interaction scenario in which the subjects were instructed to interact with an affectively colored artificial agent. Audio and video material was collected from 13 different subjects in overall 63 recordings. The recorded data was labeled in four affective dimensions: *arousal*, *expectancy*, *power* and *valence*. The annotations of the raters have been averaged for each dimension, resulting in a real value for each time step. Subsequently, the labels are binarized using a threshold equal to the grand mean of each dimension. Two to eight raters annotated every recording. Along with the sensor data and the annotations, a word-by-word transcription of the spoken language was provided which partitions the dialog into conversational turns. For the evaluation of the challenge, only arousal was taken into account as classification of the other dimensions yielded poor results.[1] (Schuller et al., 2011); McKeown et al., 2010) for a detailed description of the data set.)

3.2 Features

In the following, the proposed method for the classification from multiple sources is described. We begin with the description of the individual modalities and extracted features.

3.2.1 Audio Features

From the audio signal, the following features have been applied:

- The *fundamental frequency* values are extracted using the f_0 tracker available in the ESPS/*waves*+[2] software package. Besides the track, the *energy* and the *linear predictive coding* (LPC) of the plain wave signal is extracted (Hermansky, 1990). All three features are concatenated to a 10-dimensional early fusion feature vector.

[1] http://sspnet.eu/avec2011/

- The *Mel frequency cepstral coefficient* (MFCC) representation is inspired by the biological known perceptual variations in the human auditory system. The perception is modeled using a filter bank with filters linearly spaced in lower frequencies and logarithmically in higher frequencies in order to capture the phonetically important characteristics of speech. The MFCCs are extracted as described in Rabiner and Juang (1993).
- The *perceptual linear predictive* (PLP) analysis is based on two perceptually and biologically motivated concepts, namely the critical bands, and the equal loudness curves. Frequencies below 1 kHz need higher sound pressure levels than the reference, and sounds between 2 and 5 kHz need less pressure, following the human perception. The critical band filtering is analogous to the MFCC triangular filtering, apart from the fact that the filters are equally spaced in the Bark scale (not the Mel scale) and the shape of the filters is not triangular, but rather trapezoidal. After the critical band analysis and equal loudness conversion, the subsequent steps required for the relative spectral (RASTA) processing extension follow the implementation recommendations in Zheng et al. (2001). After transforming the spectrum to the logarithmic domain and the application of RASTA filtering, the signal is transformed back using the exponential function.

3.2.2 Video Features

We investigated a biologically inspired model architecture to study the performance of form and motion feature processing for emotion classification from facial expressions. The model architecture builds upon the functional segregation of form and motion processing in primate visual cortex. Initial processing is organized along two mainly independent pathways, each specialized for the processing of form as well as motion information, respectively.

We have directly utilized the two independent data streams for visual analysis of facial features in (Glodek et al., 2011) and already achieved robust results for automatic estimation of emotional user states from video only and audio-visual data. Here, we extended the basic architecture by further subdividing the motion-processing channel. We argue that different types of spatio-temporal information are available in the motion representation, which can be utilized for robust analysis of facial data. On the *global scale* the overall, external,

[2] http://www.speech.kth.se/software/

motion of the face is indicative for pose changes and non-verbal communication signals, e.g., head movements during nodding or selective shifts of attention through pose changes. On the other hand,on the local scale, the internal facial motions are indicative of fine-grained changes in the facial expression and emotional exposition. Examples are, e.g., eye blinks, smiles or mouth openings. We reasoned that the segregation of this information should be helpful to further improve the analysis of emotion data and, thus, process the visual input stream along *three* independent pathways. In order to make use of more detailed task-related information, we propose here an extended model architecture which aims at first segregating form and motion, as briefly outlined above, and further subdivides the motion stream into separate representations of global and local motion, respectively. An overview of the outline of the architecture is presented in Figure 7. Motion and form features are processed along two separate pathways, composed of alternating layers of filtering (S) and non-linear pooling (C) stages. In layer S1, different scale representations of the input image are convolved with 2D Gabor filters of different orientations (form path) and a spatio-temporal correlation detector is used to build a discrete velocity space representation (motion path). The initial motion representation is then further subdivided to build separate representations of global and local facial motion. Global motion is approximated by the best-fit affine motion. To achieve this, the facial region is detected by searching for horizontally oriented barcode like structures within a Gabor-filtered input image (Dakin and Watt, 2009) which is refined into facial regions-of-interests around eyes, nose and mouth. These regions are excluded from the successive random sampling process used for the estimation of the affine transformation parameters representing the global flow (affine flow).The residual or local flow is then calculated by subtracting the affine flow from the unmodified flow to provide the input representation for extracting local motion responses. All three streams, or channels, are then further processed in parallel by hierarchical stages of alternating S- and C-filtering steps. Layer C1 cells pool the activities of S1 cells of the same orientation (direction) over a small local neighborhood and two neighboring scales and speeds, respectively. The layer S2 is created by a simple template matching of patches of C1 activities against a number of prototype patches. These prototypes are randomly selected during the learning stage (for details, see (Mutch and Lowe, 2008). In the final layer C2, the S2 prototype responses are again pooled over a limited neighborhood and combined into a single feature vector which serves as input to the successive classification stage.

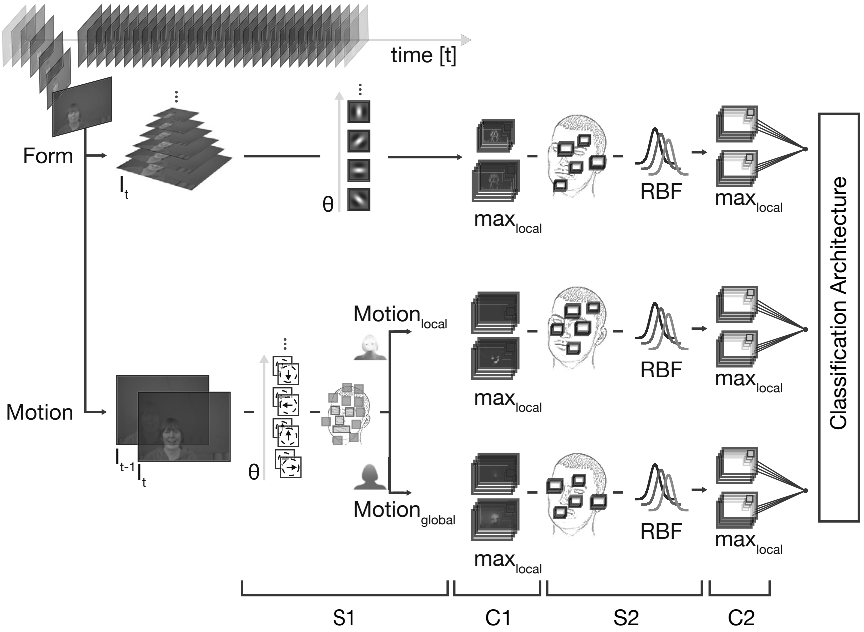

Figure 7. *Visual feature extraction.* Motion and form features are processed along two separate pathways, one form- and one motion pathway. The initial motion representation is further subdivided to build separate representations of global and local facial motion. All three streams, or channels, are further processed in parallel by hierarchical stages of alternating S- and C-filtering stepsand finally combined into a single feature vector which serves as input to the successive classification stage.

(Color image of this figure appears in the color plate section at the end of the book.)

Figure 8 demonstrates the capability of the approach to analyze differential features in non-verbal communication represented in segregated channels of visual motion information. In the case shown, the subject moves the head to point out disagreement or even disgust. This expressive communicative feature is encoded in the *global affine flow pattern* showing head motion to the right (color coded in accordance to the color map in Figure 8, right). The *local motion* activity overall depicts a brief moment in which the person opens her eyes (upward motion of eye lids (color code)) and also the chin region moves left-downwards while closing the mouth. Both motion features are now available to feed forward into the emotion classifier network for analyzing the motion related non-verbal communication behavior. Notice that in the residual flow pattern overall motion is reduced and solely local motion that is caused by facial expression remains (in the shown example caused by eye-, mouth- and cheek-movement).

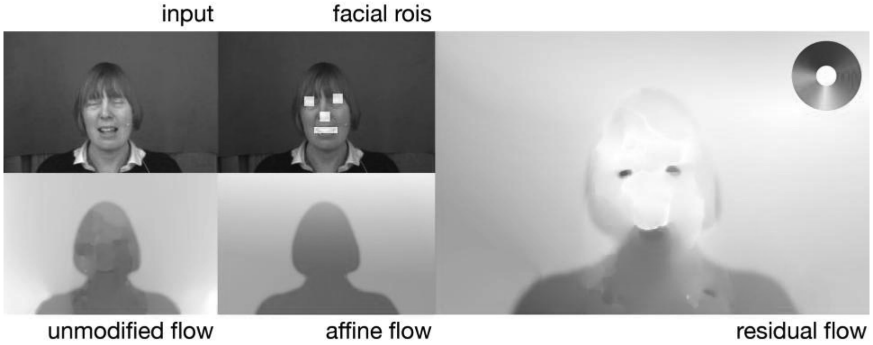

Figure 8. *Segregation of motion into separate channels of global and local motion. Four small images, in reading order:* A still image from the input sequence (input), localized eye-nose-mouth regions (facial regions of interest), optical flow calculated between two successive frames (unmodified flow) and estimation of the affine flow transformation parameters representing the global flow (affine flow).*Right image:* The residual, or local, flow is calculated by subtracting the affine flow from the unmodified flow, which reduces the overall motion energy. Residual flow caused by facial expressionremains (in theexample caused by eye-, mouth- and cheek-movement).

(Color image of this figure appears in the color plate section at the end of the book.)

3.2.3 Physiological Features

The physiological signals were acquired using a NEXUS-32 polygraph, a flexible 32 channel monitoring system. Three physiological channels were recorded: the electromyogram (EMG) of the corrugator supercilii and zygomaticus major, and the skin conductance (SCL).

To measure the SCL, two electrodes of the sensor are positioned on the index finger and the ring finger. Since the sweat glands are innervated exclusively sympathetically, i.e. without influence of the parasympathetic nervous system, the electrodermal activity is considered a good indicator of the "inner" tension of a person. This aspect can be reproduced particularly impressively by the observation of a rapid increase in skin conductance within one to three seconds due to a simple stress stimulus (e.g. deep breathing, emotional excitement or mental activity).

Electrical muscle activity is also an indicator of general psychophysiological arousal, as increased muscle tone is associated with increasing activity of the sympathetic nervous system, while a decrease in somatomotor activity is associated with predominantly parasympathetic arousal. We used two channel EMGs for corrugator and zygomaticus muscles. EMG responses over facial muscle regions like corrugator supercilii, which draws the brow downward and medial ward to form a frown, and zygomaticus major, which elevates the corner of the mouth superiorly and posteriorly to produce a smile, can effectively discriminate valance (pleasure) and intensity of emotional states.

In general, a slow low- or band-pass filter is applied together with a linear piecewise detrending of the time series at a 10-s basis. From the subject's respiration, the following features (Boiten et al., 1994) are computed (low-pass filtered at 0.15 Hz): mean and standard deviation of the first derivatives (10-s time window), breathing volume, mean and standard deviation of breathe intervals and Poincaré plots (30-s time window each). The EMG signals were used to compute the following features (band-pass filtered at 20–120 Hz, piecewise linear detrend): mean of first and second derivatives (5-s time window) (Picard, 2003) and power spectrum density estimation (15-s time window) (Welch, 1967). The following features are extracted from the skin conductance (SCL) (low-pass filtered at 0.2 Hz): mean and standard deviation of first and second derivative (5-s time window).

3.3 Statistical evaluation

In this section, the statistical evaluation of the architecture is described for the mentioned data. All reported results originate from leave one subject out experiments.

3.3.1 EmoRec II

Classification of Spoken Utterances

The MFCC features have been calculated using 40-ms windows and were averaged to form 200-ms blocks such that all features have a uniform alignment. The three available features were separately classified using a SVM with an RBF kernel function and a probabilistic output function. For the individual features, accuracies from 52.5% to 57.9% were accomplished. Furthermore, these results were combined using the average of the confidence values and a temporal fusion of 10 s was conducted. This resulted in an accuracy of 55.4% (compare Table 1).

Classification of Facial Expressions

We classified the facial expressions using a multivariate Gaussian. In order to render a stable classification, a bagging procedure was conducted and a reject option was implemented: hereby 99% of the test frames were rejected with respect to the confidence of the classifier. An accuracy of 54.5% was achieved and only 52.3% without reject option (see Table 1).

Table 1. Accuracy of each unimodal classifier. Results in percent with standard deviation.

Feature	Accuracy	Accuracy (reject)
MFCC 40 ms	57.1% (2.4%)	n/a
MFCC 200 ms	57.9% (2.1%)	n/a
ModSpec	52.5% (2.2%)	n/a
Fusion of Audio channel	55.4% (4.6%)	n/a
EMG (5 s)	56.5% (10.7%)	69.5% (14.0%)
EMG (20 s)	50.2% (11.0%)	52.4% (24.6%)
SCL (5 s)	52.8% (4.5%)	44.1% (3.4%)
Respiration (20 s)	52.6% (8.4%)	44.6% (17.0%)
Fusion of physiology channel	55.6% (14.5%)	59.7% (13.0%)
Facial Expression	52.3% (4.4%)	54.5% (4.5%)

Classification of Biophysiological Signals

The described features were partitioned into different sets defined by the feature type and window size. This results in four different feature sets. Each of the sets was classified by a perceptron using bagging. Furthermore, a reject option of 80% is used. The setting results in accuracies ranging from 69.5% to 50.2%. Without reject option, the accuracies were lower.

These intermediate results were combined based on a new superordinate time window of 60 s with an offset of 30 s: first the confidences of the individual channels are averaged followed by the combination of the channels, resulting in accuracies of 59.7% (55.6% without reject; compare Table 1).

Multi-modal Combination

The intermediate results of the unimodal classifiers are combined to obtain the final class membership. The final time granularity was set to 60 s to evaluate all combinations of modalities. If no decision can be made (e.g. due to rejection of samples), the overall window is rejected. The weights for the classifiers have been chosen equally for all classifier combinations. An additional study, putting a weighting with focus on audio and physiology according to the respective performance, was conducted. The results are shown in Table 2 and vary around an accuracy of 60%. The highest accuracy is achieved using solely audio and video. Generally, the standard deviations are high. When all available sources are combined, the mean accuracy slightly drops, but on the other hand, the standard deviation decreases. Further details can be found in Schels et al. (2012).

Table 2. Accuracies of every multi-modal combination. Results in percent with standard deviation.

Combination	Accuracy
Audio (1) + Video (1)	62.0% (15%)
Video (1) + Physiology (1)	59.7% (13.4%)
Audio (1) + Physiology (1)	60.2% (9.3%)
Video (1) + Audio (1) + Physiology (1)	60.8% (9.1%)
Video (1) + Audio (2) + Physiology (3)	61.5% (12.6%)

3.3.2 AVEC 2011 Data Collection

For each label dimension and for each audio feature, a bag of hidden Markov models (HMMs) have been trained (Breimann, 1996; Rabiner, 1989). The hidden states and the number of mixture components of the HMM have been optimized using a parameter search, resulting in the selection of three hidden states and two mixture components in the Gaussian mixture model (GMM) having full covariance matrices.

The evaluation of the optimization process further inferred that some features appear to be inappropriate to detect certain labels. It turned out that only the label *arousal* can draw information from all features, *expectancy* and *power* performed better using only the energy, fundamental frequency and the MFCC features. The label *valance* favored only the MFCC features. For each label, the log-likelihoods of every HMM, trained on the features, are summed. To obtain more robust models, we decided to additionally use five times as many models per class and summed the outcome as well.

Furthermore, the assumption was made that the labels are changing only slowly over time. We therefore conducted the classification on turn basis by collecting the detections within one turn and multiplied the likelihoods to obtain more robust detections. A schema visualizing the applied fusion architecture is shown in Figure 9. The results of this approach are reported in Table 3.

Within the video challenge, the n-SVM was employed as base classifier (Schölkopf et al., 2000). The implementation was taken from the well-known LibSVM repository. We concatenated 300 form and 300 motion features and used them to train a n-SVM using a linear kernel and probabilistic outputs according to Platt (1999). Due to memory constraints, only 10.000 randomly drawn samples were used.

Again a parameter search was applied to obtain suitable parameters, resulting in setting $n = 0.3$ for *arousal* and *power* and $n = 0.7$ for *expectancy* and *valence*. Based on the results of all label dimensions,

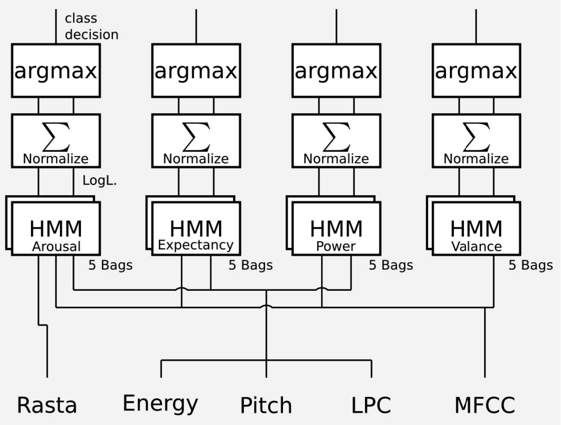

Figure 9. Architecture of the audio classifier. For each label, a bag of HMMs have been trained on selected feature sets.

Table 3. Classification results of the AVEC 2011 development data set. The weighted accuracy (WA) corresponds to the correctly detected samples divided by the total number of samples. The unweighted accuracy (UA) is given by the averaged recall of the two classes.

	Arousal		**Expectancy**		**Power**		**Valence**	
	WA	UA	WA	UA	WA	UA	WA	UA
Audio	66.9	67.5	62.9	58.5	63.2	58.4	65.7	63.3
Visual	58.2	53.5	53.5	53.2	53.7	53.8	53.2	49.8
Audio/Visual	69.3	70.6	61.7	60.1	61.3	59.1	68.8	66.4

an intermediate fusion was conducted using an MLP to obtain the final prediction. A schema illustrating the architecture used is shown in Figure 10. The results are reported in Table 3.

Considering the audiovisual challenge, we used the same approach for each modality as described in the earlier sections but omitted the last layer in which the class decision was performed. The probabilistic outputs of the video stream are collected using averaging and multiplication with a subsequent normalization such that the decisions are on word level. The HMM log-likelihoods of the label dimensions are transformed and normalized such that they are ranging between zero and one.

By concatenating the results of all label dimensions, a new 12-dimensional feature vector is obtained. The new features are then used to train an intermediate fusion layer based on an MLP.

Like in the audio challenge, the final decision is done on a turn basis by collecting the outputs within one turn and fusing them using

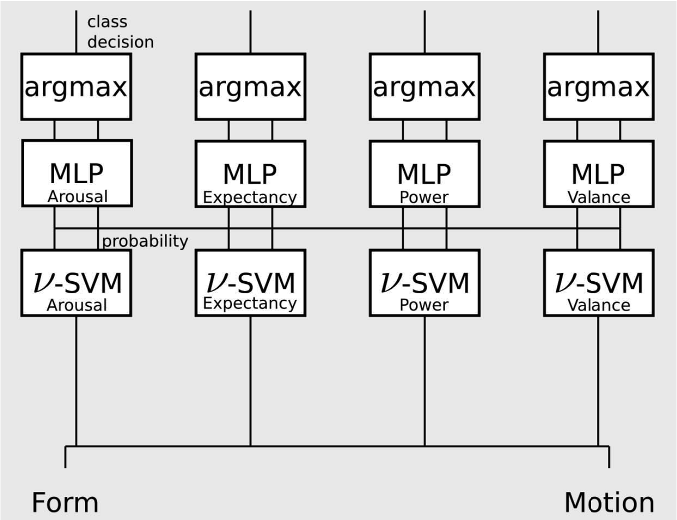

Figure 10. For the video-based classification, form and motion features are concatenated and used to train *n*-SVM for each label dimension. The outputs of the classifiers are used to train an intermediate fusion layer realized by MLPs.

multiplication. Figure 11 shows the audiovisual classifier system, while the results are given in Table 3.

4. Conclusion and Future Work

Classifying the emotion is generally a difficult task when leaping from overacted data to realistic human-computer interaction. In this study, the problem was investigated by combining different modalities. The result of the evaluation shows that the usage of different modalities

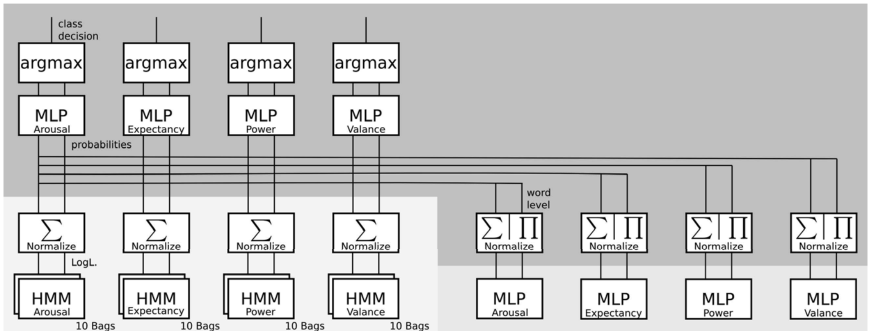

Figure 11. Overall architecture of the audiovisual classifier system, the outputs of all modalities are integrated on word level and used to train a multilayer neural network for each label dimension.

can reduce the testing error. On the other hand, the variances of the classification are relatively high.

Rejecting samples when classifying such kind of data turns out to be a sound approach leading to more robust results, especially when the distribution of the classes in the data is heavily overlapping. In future work, it could be promising to implement an iterative classifier training procedure, where the training data can be rejected.

The results presented in Tables 1 to 3 are preliminary and must be further evaluated in several directions:

1. Feature extraction techniques as described in the previous sections have been successfully applied to the recognition of Ekman's six basic emotions for benchmark data sets consisting of acted emotional data. In these data sets, emotions shown by the actors are usually over-expressed and different from the emotional states that can be observed in the AVEC data set.
2. The classifier architecture is based on the so-called late fusion paradigm. This is a widely used fusion scheme that can be implemented easily just by integrating results of the pre-trained classifier ensemble by fixed or trainable fusion mappings, but more complex spatio-temporal patterns on an intermediate feature level cannot be modeled by such decision level fusion scheme.
3. The emotional states of the AVEC2011 data set are encoded by crisp binary labels, but human annotators have usually problems to assign a confident crisp label to an emotional scene (e.g. single spoken word or a few video frames) or disagree, and thus dealing with fuzzy labels or labels together with a confidence value during annotation and classifier training phase could improve the overall recognition performance.

Acknowledgments

This chapter is based on work done within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). Tobias Brosch and Miriam Schmidt are supported by scholarships of the graduate school Mathematical Analysis of Evolution, Information and Complexity of the University of Ulm.

REFERENCES

Abdel Hady, M. and F. Schwenker. 2010. Combining committee-based semi-supervised learning and active learning. *Journal of Computer Science and Technology,* **25(4):**681–698.

Argyle, M. 1988. Bodily Communication. Routledge, London (GB).

Bayerl, P. and H. Neumann. 2007. A fast biologically inspired algorithm for recurrent motion estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* **29(2):**246–260.

Bishop, C. 2006. Pattern Recognition and Machine Learning. Springer, New York (NY).

Black, M. and Y. Yacoob. 1997. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision,* **25:**23–48.

Blake, A., B. Bascle, M. Isard, and J. MacCormick. 1998. Statistical models of visual shape and motion. *Phil. Trans. Royal Society of London, Series A,* **386:**1283–1302.

Boiten, F., N. Frijda and C. Wientjes. 1994. Emotions and respiratory patterns: Review and critical analysis. *International Journal of Psychophysiology,* **17(2):**103–128.

Breiman, L. 1996. Bagging predictors. *Journal of Machine Learning,* **24(2):**123–140.

Cannon, W. 1927. The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology,* **39(1/4):**106–124.

Caridakis, G., G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis and K. Karpouzis. 2007. Multimodal emotion recognition from expressive faces, body gestures and speech. In Artificial Intelligence and Innovations, from Theory to Applications, Vol. 247, pp. 375–388. Springer.

Chen, L. and T. Huang. 2000. Emotional expressions in audiovisual human computer interaction. *IEEE,* **1:**423–426.

Cootes, T. and C. Taylor. 1992. Active chape models—"smart snakes". Proceedings of the British Machine Vision Conference (BMVC'92), pp. 266–275.

Cootes, T., G. Edwards and C. Taylor. 2001. Acttive appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* **23(6):**681–685.

Dakin, S. and R. Watt. 2009. Biological "bar codes" in human faces. *Journal of Vision,* **9(4)(2):**1–10.

Dietrich, C., G. Palm and F. Schwenker. 2003. Decision templates for the classification of bioacoustic time series. *Journal of Information Fusion,* **4(2):**101–109.

Dietrich, C., F. Schwenker and G. Palm. 2001. Classification of time series utilizing temporal and decision fusion. Workshop on Multiple Classifier Systems. LNCS 2096, pp. 378–387. Springer.

Ekman, P. and W. Friesen. 1978. Facial Action Coding System: Investigator's Guide. Consulting Psychologists Press, Palo Alto (CA).

Eyben, F., M. Wöllmer and B. Schuller, 2012. A multitask approach to continuous five-dimensional affect sensing in natural speech. *ACM Transactions on*

Interactive Intelligent Systems (TiiS), **2(1):**6.

Fischler, M. and R. Bolles. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM,* **24(6):**381–395.

Glodek, M., L. Bigalke, M. Schels and F. Schwenker. 2011. Incorporating Uncertainty in a Layered HMM architecture for Human Activity Recognition. Proceedings of the Joint Workshop on Human Gesture and Behavior Understanding (J-HGBU), pp. 33–34, ACM.

Glodek, M., G. Layher, F. Schwenker and G. Palm. 2012. Recognizing Human Activities using a Layered Markov Architecture. Proceeding of the International Conference on Artificial Neural Networks (ICANN. LNCS 7552, pp. 677–684. Springer.

Glodek, M., M. Schels, G. Palm and F. Schwenker. 2012. Multi-Modal Fusion Based on Classification Using Rejection Option and Markov Fusion Network. Proceedings of the International Conference on Pattern Recognition ICPR, pp. 1084–1087, IEEE.

Glodek, M., M. Schels, G. Palm and F. Schwenker. 2012. Multiple Classifier Combination Using Reject Options and Markov Fusion Networks. Proceedings of the International Conference on Multimodal Interaction, pp. 465–472, ACM.

Glodek, M., S. Scherer, F. Schwenker. 2011. Conditioned Hidden Markov Model Fusion for Multimodal Classification. Proceedings of the 12th European Conference on Speech Communication and Technology (Interspeech'11), pp. 2269–2272. ISCA.

Glodek, M., S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm and F. Schwenker. 2011. Multiple Classifier Systems for the Classification of Audio-Visual Emotional States. Proceedings of the International Audio/Visual Emotion Challenge (AVEC) and Workshop. LNCS 6975, pp. 359–368, Springer.

Haykin, S. 1999. Neural Networks: A Comprehensive Foundation. Prentice Hall, East Lansing (MI).

Hermansky, H. 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America,* **87(4):**1738–1752.

Isard, M. and A. Blake. 1998. CONDENSATION—Conditional density propagation for visual ttracking. *International Journal of Computer Vision,* **29:**5–28.

Kelley, J. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems,* **2(1):**26–41.

Kuncheva, L.I. 2004. Combining Pattern Classifiers: Methods and Algorithms. Wiley & Sons, Hoboken (NJ).

Layher, G., S. Tschechne, S. Scherer, T. Brosch, C. Curio and H. Neumann. 2011. Social Signal Processing in Companion Systems—Challenges Ahead. Proceedings of the Workshop Companion-Systeme und Mensch-Companion Interaktion (41st Jahrestagung der Gesellschaft für Informatik).

McKeown, G., M. Valstar, R. Cowie and M. Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. *IEEE*, pp. 1079–1084.

Meudt, S., L. Bigalke and F. Schwenker. 2012. ATLAS—An Annotation Tool for HCI Data Utilizing Machine Learning Methods. Proceedings of the 1st International Conference on Affective and Pleasurable Design (APD'12) [Jointly with the 4th International Conference on Applied Human Factors and Ergonomics (AHFE'12)], pp. 5347–5352, CRC Press.

Mutch, J. and D. Lowe. 2008. Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision,* **80(1):**45–57.

Oudeyer, P. 2003. The production and recognition of emotions in speech: Features and algorithms. *International Journal of Human-Computer Studies,* **59(1–2):**157–183.

Picard, R. 2003. Affective computing: Challenges. *International Journal of Human-Computer Studies,* **59(1):**55–64.

Platt, J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers, pp. 61–74.

Poggio, T., U. Knoblich and J. Mutch. 2010. CNS: A GPU-based framework for simulating cortically-organized networks. *MIT-CSAIL-TR-2010-013/ CBCL-286.*

Rabiner, L. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE,* **77(2):**257–286.

Rabiner, L. and B. Juang. 1993. Fundamentals of speech recognition. Prentice-Hall Signal Processing Series, East Lansing (MI).

Rolls, E. 1994. Brain mechanisms for invariant visual recognition and learning. *Behavioural Processes,* **33(1–2):**113–138.

Schölkopf, B., A. Smola, R. Williamson and P. Bartlett. 2000. New support vector algorithms. *Neural Computation,* **12(5):**1207–1245.

Schachter, S. 1964. The interaction of cognitive and physiological determinants of emotional state. In L. Berkowitz (ed.), Advances in Experimental Social Psychology (Vol. 1, pp. 49–80). Academic Press.

Schels, M. and F. Schwenker. 2010. A multiple classifier system approach for facial expressions in image sequences utilizing gmm supervectors. Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10), pp. 4251–4254. IEEE.

Schels, M., M. Glodek, S. Meudt, M. Schmidt, D. Hrabal, R. Böck, S. Walter, F. Schwenker. 2012. Multi-Modal Classifier-Fusion for the Classification of Emotional States in WOZ Scenarios. Proceedings of the 1st International Conference on Affective and Pleasurable Design, pp. 5337–5346. CRC Press.

Schels, M., S. Scherer, M. Glodek, H.A. Kestler, G. Palm and F. Schwenker. 2011. On the discovery of events in EEG data utilizing information fusion. *Journal on Multimodal User Interfaces*, **6(3-4):**117–141.

Scherer, S., M. Glodek, G. Layher, M. Schels, M. Schmidt, T. Brosch, S. Tschechne, F. Schwenker, H. Neumann, G. Palm. 2012. A generic framework for the inference of user states in human computer interaction: How patterns of low level communicational cues support complex affective states. *Journal on Multimodal User Interfaces*, **2(1):**4:**1–4:**31.

Scherer, S., M. Glodek, F. Schwenker, N. Campbell and G. Palm. 2012. Spotting laughter in natural multiparty conversations: a comparison of automatic online and offline approaches using audiovisual data. *ACM Transactions on Interactive Intelligent Systems,* **2(1):**1–31.

Schuller, B., G. Rigoll and M. Lang. 2003. Hidden markov model-based speech emotion recognition. *IEEE*, **2:**401–404.

Schuller, B., M. Valstar, F. Eyben, G. McKeown, R. Cowie and M. Pantic. 2011. The First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011).

Schwenker, F. and E. Trentin. 2012. Proceedings of the First IAPR TC3 Workshop on Partially Supervised Learning, LNCS 7081. Springer.

Sebe, N., M.S. Lew, Y. Sun, I. Cohen, T. Gevers and T.S. Huang. 2007. Authentic facial expression analysis. *Image Vision Comput.,* **25(12):**1856–1863.

Serre, T. and M.A. Giese. 2011. Elements for a neural theory of the processing of dynamic faces. In C. Curio, H. Bülthoff and M.A. Giese (eds.), Dynamic Faces. Insights from Experiments and Computation, pp. 187–210. MIT Press.

Settles, B. 2009. Active Learning Literature Survey. Department of Computer Sciences, University of Wisconsin-Madison.

Thiel, C., F. Schwenker and G. Palm. 2005. Using Dempster-Shafer Theory in MCF Systems to Reject Samples. Proceedings of the 6th International Workshop on Multiple Classifier System, LNCS 3541, pp. 118–127. Springer.

Walter, S., S. Scherer, M. Schels, M. Glodek, D. Hrabal, M. Schmidt, R. Böck, K. Limbrecht, H.C. Traue, F. Schwenker. 2011. Multimodal emotion classification in naturalistic user behavior. In Human-Computer Interaction, Part III, HCII 2011, J. Jacko (ed.), LNCS 6763, pp. 603–611. Springer, Berlin (DE).

Welch, P. 1967. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio and Electroacoustics,* **15(2):**70–73.

Zeng, Z., M. Pantic, G. Roisman and T. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* **31(1):**39–58.

Zheng, F., G. Zhang and Z. Song. 2001. Comparison of different implementations of MFCC. *Journal of Computer Science and Technology,* **16(6):**582–589.