Eurographics/SIGGRAPH Symposium on Computer Animation (2003)
D. Breen, M. Lin (Editors)

Unsupervised Learning for Speech Motion Editing

Yong Cao[1,2] Petros Faloutsos[1] Frédéric Pighin[2]

[1]University of California at Los Angeles, Department of Computer Science
[2]University of Southern California, Institute for Creative Technologies

Abstract

We present a new method for editing speech related facial motions. Our method uses an unsupervised learning technique, Independent Component Analysis (ICA), to extract a set of meaningful parameters without any annotation of the data. With ICA, we are able to solve a blind source separation problem and describe the original data as a linear combination of two sources. One source captures content (speech) and the other captures style (emotion). By manipulating the independent components we can edit the motions in intuitive ways.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Animation, I.5.1 [Pattern Recognition]: Statistical

1. Introduction

In this paper we address the problem of editing recorded facial motion. Editing motion capture data is an important problem, since without the ability to edit recorded motion we can only replay the recorded data.

Producing high quality facial animation is one of the most challenging problems in computer animation. Hundreds of individual muscles contribute to the generation of complex facial expressions and speech. The dynamics of these muscles are not well understood and no system to date can simulate realistic faces in real-time. Using motion capture is an attractive alternative. Current motion capture technology can record accurately the motions of a subject's face. These motions can then be mapped onto a face model to produce realistic animations. It is however impossible to record all the motions a face can do. Thus, the motion capture sessions should be carefully planned to meet the needs of the production. A different approach is to record a representative set of motions and use machine learning techniques to estimate a generative statistical model. The goal is then to find and fit a model that is able to resynthesize the recorded data. Finding an appropriate model that can reproduce the subtleties of the recorded motion can be a very difficult task. In addition, the parameters of the model might not be appropriate for manipulating or editing the data. Fitting statistical models generally involves minimizing an error function regardless of the semantics of the data or what the model's parameters really represent. Interpretation of the data is generally best done by a human observer that can annotate the data and specify its semantics. A function can then be learnt that expresses the correlation between the annotations, the input, and the motions, the output. We can use this function to manipulate the data. However, when the size of the data becomes large, human intervention and annotation is impractical. Our method addressed this issue.

In this paper we propose an unsupervised learning technique, based on *Independent Component Analysis* (ICA), that splits the recorded motions into linear mixtures of statistically independent sources. These sources, called independent components, offer a compact representation of the data with clear semantics. The lack of structure or model underlying the recorded data makes it really hard to edit. In contrast, the decomposition we propose provides a meaningful parameterization of the original data that is suitable for editing. The technique is automatic and does not require annotating the data.

The remainder of the paper is organized as follows. Section 2 reviews related research. Section 3 introduces ICA and describes its application to recorded facial motion. Section 4 explains how we determine the semantics of the resulting decomposition. Section 5 describes editing operations using the ICA representation of the motion. Section 6 presents our experimental results. Lastly, Section 7 summarizes our approach and proposes future directions.

2. Related work

Two bodies of research are germane to our study: face motion synthesis and motion analysis. We briefly describe previous work in these two fields and position our approach.

2.1. Face motion synthesis

The motions of the human face can be simulated using a physically based model. Physics-based approaches use simulated muscle models to deform a three-dimensional face mesh. Despite their computational expense they has been shown to be quite effective [16, 26].

Most speech animation systems exploit the fact that speech can be reliably segmented into units (e.g. phonemes). The voice track is manually [21] or automatically [17, 6] segmented into phonemic representations which are then mapped to lips shapes. Of particular importance here is the problem of *co-articulation*. Co-articulation means that the mouth shape used to produce a particular phoneme depends not only on the current phoneme but also on the phoneme before and after the current one. *Hidden Markov Models* (HMM) have been used extensively to represent transitions between phonemic representations with proper co-articulation.

The synthesis of speech animation requires a library of lips shapes that can be matched with speech units. This library can be designed in several ways. One option is to create manually each shape. For realistic animation it however preferable to record these shapes using video or motion capture data. *Video Rewrite* [6] is a representative example of such techniques. It constructs a large database of audiovisual basis units, based on triphones. Given a novel input utterance, the corresponding facial motion is constructed by concatenating the appropriate triphones from the database. In order to be useful the method requires a large database of triphones, which leads to a scaling problem. To eliminate the need for large example databases, a statistical face motion can be estimated from the data. *Voice Puppetry* [4] develops a mapping from voice to face by learning a model of a face's observed dynamics. The model takes into account the position and the velocity of facial features and learns a probability distribution over the different facial configurations. Ezzat et al [12] develop a *Multidimensional Morphable Model* for the voice to face mapping focusing on lip-syncing. Head and upper face motion is dealt with in an ad hoc fashion.

While the previous techniques can generate high quality motion they generally do not provide the animator with intuitive control over the emotional state of the talking face. They focus on the mapping of the audio and visual speech signal and effects such as co-articulation. In contrast our work develops an unsupervised learning approach that learns two separate mappings, one between the phonemic content of the audio signal and the motion of the face and another between the audio signal and the emotional content of the speech.

2.2. Motion analysis

Motion capture allows the recording a high fidelity motions from live actors. This technique spurred a wealth of research efforts in motion analysis.

Chuang et al [7] present an interesting attempt to separate visual speech into content and style (emotion). Their method based on factorization [13, 24] produces a bilinear model that extracts emotion and content from input video sequences. However, their approach normalizes the signals losing important temporal information and it is tailored to video data. It is not clear whether it would transfer to 3D.

The pattern recognition community has performed a significant amount of work on facial expression analysis. Expressions are typically based on tracking the motion of particular facial elements such as the eyes, the rigid body motion of the face, or transient features such as wrinkles [10, 9, 11, 8, 3, 18, 2]. These systems are quite effective for recognition, however, it is not clear how they can be used to synthesize or edit facial motion.

Learning the style and the content from recorded variations of a motion has been also investigated in the area of full body animation. Pullen et al [22] propose a technique that decomposes motion into different frequency bands. The low frequency components represent the basic motion while the higher frequency ones capture the style of the motion. Combining the basic signals with different higher frequency bands results into stylistic variations of the basic motions. Brand et al [5] train *Hidden Markov Models* to capture the style variations of example dance data. The resulting style models can be applied to novel dance sequences. Unuma et al [25] decompose example motion into high and low frequencies using Fourier analysis. Manipulating the resulting coefficients provides an intuitive way to alter the original motion.

Our work employs *Independent Component Analysis* to extract style and content models from a large set of recorded facial motions. The resulting independent components are the basis for an intuitive visual speech editing tool. Mori et al [19] proposes a motion editing and synthesis technique for human walking based on *Independent Component Analysis*. However, the paper is too short to allow us to evaluate the results.

3. Facial motion decomposition

In this section we present an overview of *Independent Component Analysis*. We then discuss our decomposition technique and the way we determine the semantics of the resulting independent components.

3.1. Independent Component Analysis

Independent Component Analysis is an unsupervised learning technique [14]. It assumes that a set of observed random variables can be expressed as linear combinations of independent latent variables. In a way it deconvolves the recorded signals into a set of statistically independent random variables. It is often associated with the "Blind Source Separation" problem. One instance of this problem can be found in audio processing: imagine that the sound in a room comes from two sources, the voice of a speaker and the humming of an air conditioning system. Solving the blind source separation problem in this context would involve recording the sound in the room (from 2 different locations) and processing it statistically so that the two original sources can be separated. This audio separation problem is a very difficult one. ICA can successfully separate the two sources by exploiting their statistical independence.

Let us examine the mathematics of ICA. Assume that we observe n random variables $x_1,\ldots,x_n$ each of them being a linear mixture of n latent or hidden variables $u_1,\ldots,u_n$, such that

$$x_j = \sum_{i=1}^{n} a_{ji} u_i,$$

or in matrix notation

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{u}. \tag{1}$$

Equation 1 represents a generative model: it describes how the recorded data $\boldsymbol{x}$ is generated by the sources $\boldsymbol{u}$. The sources u_i, which are called the independent components, cannot be observed directly. The matrix of coefficients $\boldsymbol{A}$, called mixing matrix, is also unknown. ICA provides a framework to estimate both $\boldsymbol{A}$ and $\boldsymbol{u}$. In practice, estimating $\boldsymbol{A}$ is sufficient, since if the matrix is known, its inverse, $\boldsymbol{W}$, can be applied to obtain the independent components:

$$\boldsymbol{u} = \boldsymbol{W}\boldsymbol{x}.$$

To estimate the matrix $\boldsymbol{A}$, ICA takes advantage of the fact that the components are statistically independent. The key to estimating the ICA model is non-gaussianity. According to the Central Limit Theorem the sum of two independent random variables usually has a distribution closer to a gaussian distribution. The idea then is to iteratively extract random variables from the recorded data that are as non-gaussian as possible. How non-gaussianity is measured is beyond the scope of this paper. Different metrics have been used, leading to a variety of implementations. For more details see [15, 14]. In our experiments we use a publicly available implementation called FastICA [1].

3.2. Preprocessing

Before we can apply ICA our data has to go through a preprocessing phase that consists of two steps, *centering* and *whitening*.

Centering shifts the data towards its mean so that the resulting random variables have zero mean. *Whitening* transforms the centered set of observed variables into a set of uncorrelated variables. *Principal Component Analysis* (PCA) can be used to perform this transformation. After preprocessing the model of Equation 1 takes the form

$$\boldsymbol{x} = E\{\boldsymbol{x}\} + \boldsymbol{P}\boldsymbol{A}\boldsymbol{u}, \tag{2}$$

where $E\{\boldsymbol{x}\}$ is the expectation of $\boldsymbol{x}$ and $\boldsymbol{P}$ is an n by m matrix obtained by applying PCA to the centered data. m is the number of principal components we keep. Matrix $\boldsymbol{P}$ will not be square $(m < n)$ if we decide to only retain a subset of the principal components. This reduction in dimension reduces the number of independent components to the same number as well.

3.3. PCA vs ICA

PCA and ICA are related statistical techniques. They both provide a linear decomposition of sampled data. The fundamental difference is that PCA assumes the latent variables are uncorrelated whereas ICA assumes they are independent. Independent random variables are also uncorrelated but not vice versa. The goal of PCA is to find a sequence of uncorrelated random variables (components) where each variable covers as much of the variance of the data as possible. The resulting sequence is ordered by decreasing variance coverage. For this reason, PCA is often an effective compression technique: by keeping the first few components most of the variance in the data can be covered. The independent components produced by ICA provide a separation mechanism between sources that are assumed independent rather than a compression mechanism.

3.4. Application to facial motion

Applying ICA to recorded facial motion is straightforward. The motion is represented as a set of time series $x_i(t)$, that captures the euclidian coordinates of the motion capture markers in time. Each of these time series can be thought of as samples of random variables x_i. Then we can directly apply ICA decomposition on this set of variables, x_i, using Equation 2.

This decomposition results into a set of independent components that have intuitive interpretation. In the next section we present how we determine the meaning of the independent components.

4. Interpretation of the independent components

Our work decomposes speech related motions into a set of sources that can be clearly interpreted and manipulated for editing purposes. In particular, we separate the data into style and content components. In our case we equate style with expressiveness or emotion and contents with the part of the motion responsible for the formation of speech.

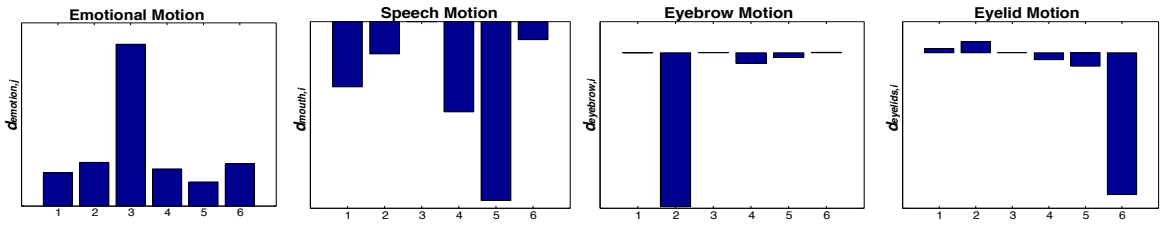

Figure 1: *These graphs illustrate the classification of independent components. Each graph illustrates a category of motion: (a) for emotions, (b) for speech, (c) for eyebrows, and (d) for eyelids. The horizontal axis represents the index of independent components. The vertical axis shows the distance metrics that we described in Section 4.*

4.1. Number of independent components

Before applying ICA on the data, we have to determine the number of components we need to extract. There is no clear rule to help us make this decision. In practice the whitening preprocessing step (Section 3) reduces the dimension of the data and determines the number of independent components. We experimentally determine how many components to keep so that we preserve the subtleties of the motion. For most of the experiments, keeping enough components to cover $95\% - 98\%$ of the variance proved to be sufficient.

In what follows we describe how we associate specific meaning to the independent components.

4.2. Emotion

We recorded the motion of an actor's face while he was uttering a set of sentences multiple times, each time expressing a different emotion. Let us denote as $(\boldsymbol{x}^i, \boldsymbol{y}^i)$, p pairs of motions that corresponds to the same sentence but two different emotions. Applying ICA to each pair of motions in our dataset, results into pairs of corresponding independent component sets, $(\boldsymbol{u}^i, \boldsymbol{v}^i)$. We would expect that the independent components related to emotion differ significantly between two speech motions that have the same content but different emotion. In contrast, if an independent component is not related to emotion, its value in time for two corresponding motions should be the same except some timing differences. In order to verify this property, we align each pair of corresponding motions using a *Dynamic Time-Warping* (DTW) algorithm[23]. Let us denote $(\boldsymbol{u}'^i, \boldsymbol{v}'^i)$ the independent components of two aligned motions after time warping. We compute their difference using the Root Mean Square (RMS) error as follows:

$$d_{emotion,j} = \left(\frac{1}{\sum q_i}\left(\sum_{i=1}^{p}\left(\sum_{k=1}^{q_i}(u'^i_j(t_k) - v'^i_j(t_k))^2\right)\right)^{\frac{1}{2}},$$

where q_i is the number of aligned time samples for pair i. The distance $d_{emotion,j}$ is designed such that it should be large if component j is related to emotion.

Figure 1(a) shows a plot of the $d_{emotion,j}$ values of 6 independent components estimated from 32 pairs of sentences of *Frustrated* and *Happy* motions. This data totals 11883 frames or 99 seconds. A clear peak can be observed for the third component. This strongly indicates that this component is related to emotional variations. The other components participate to a lesser degree to the emotional content of the motions. This shows that speech motion cannot be strictly separated into statistically independent components. Our approach is albeit a successful approximation. As further proof, in Figure 2 we plot the evolution of the different components over time for a set of five pairs of motions. On the timeline, we alternate *Frustrated* and *Happy* motions. The behavior of the third component appears very much related to changes in emotions (illustrated with different gray levels).

4.3. Content

We define content as the part of the motion associated with the formation of speech independent of expressiveness. For this case we only consider the motion of the markers in the mouth area (12 markers in our dataset).

Let us define a distance metric between two motions that have been reconstructed using two subsets of independent components, A and B.

$$d_{mouth}(\boldsymbol{x}_A, \boldsymbol{x}_B) = \left(\frac{1}{q}\sum_{k=1}^{q}\left(\frac{1}{r}\sum_{l=1}^{r}(x^l_A(t_k) - x^l_B(t_k))^2\right)\right)^{\frac{1}{2}}, \quad (3)$$

where $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are the motions reconstructed using component subset A and B respectively, q is the number of time samples of both motions, r is the number of the markers considered for the mouth region (12 markers).

Reconstructing the motion of the mouth markers using all the independent components produces $\boldsymbol{x}_{all}$. In general this is different from the captured motion because of the compression done in the preprocessing step (Section 3). In order to evaluate how much independent component i contributes to the mouth motion we compute the following metric

$$d_{mouth,i} = d_{mouth}(\boldsymbol{x}_{E\cup\{i\}}, \boldsymbol{x}_{all}) - d_{mouth}(\boldsymbol{x}_E, \boldsymbol{x}_{all}), \quad (4)$$

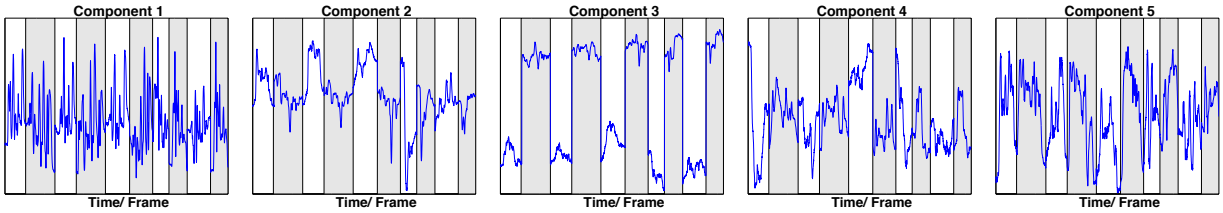

Figure 2: *These graphs present the evolution in time of five independent components corresponding to five pairs of Frustrated and Happy motions. The timeline alters between Frustrated (light) and Happy (dark) motions. Notice how the third component oscillates between extreme values when the emotion changes.*

where E is the subset of independent components responsible for emotion and $\boldsymbol{x}_E$ is the marker motion reconstructed from subset E.

In Equation 4 $d_{mouth,i}$ quantifies the influence of independent component i on the motion of the mouth. The larger in absolute value this number is, the more influence component i has over the mouth motion. Figure 1(b) shows the value of $d_{emotion,i}$ for six independent components. Notice how large $d_{mouth,1}$, $d_{mouth,4}$, and $d_{mouth,5}$ are compared to the rest of the components. We can visually verify that the motion $x_{\{1\}\cup\{4\}\cup\{5\}}$ reconstructed using components 1, 4 and 5 captures most of the speech motion.

4.4. Blinking and non-emotional eyebrow motions

Our experiments show that some independent components cannot be associated with emotion or content. We have experimentally determined that we can further classify such components into two groups: one for blinking motion and the other for non-emotional eyebrow motion. The later refers to eyebrow motion that reflects stress and emphasis in the speech rather than the emotional state of the speaker.

In order to identify the components related to these two types of motion we use the same method employed for finding content related components. We define $d_{eyebrow}$ and $d_{eyelids}$ according to Equation 3 while considering only the markers on the eyebrows and the eyelids respectively. We use these two metrics to define $d_{eyebrow,i}$ and $d_{eyelids,i}$ from Equation 4 for the eyebrows and the eyelids respectively.

Figure 1(c) shows the value of the distance metric $d_{eyebrow,i}$ for six independent components. Notice how much larger $d_{eyebrow,2}$ is compared to the distance metric of the rest of the components. Clearly component 2 captures most of the eyebrow motion. Similarly, Figure 1(d) shows the value of the distance metric $d_{eyelids,i}$ for each of the six components. In this case, $d_{eyelids,6}$ dominates the rest of the components. We conclude that component 6 captures most of the eyelid motion.

5. Editing

Based on the proposed decomposition we have built a facial motion editing tool that allows the user to interactively change the apparent emotional content of visual speech.

We have implemented multiple operations in ICA space to change the emotion expressed in a recorded motion. We use the ICA representation to resynthesize the motion after editing the parameters of the model. The ICA model can be written (see Section 3):

$$\boldsymbol{x} = E\{\boldsymbol{x}\} + \boldsymbol{PAu}. \tag{5}$$

There are three parameters that can be manipulated: the mean $E\{\boldsymbol{x}\}$, the mixing matrix $\boldsymbol{PA}$, and the independent components $\boldsymbol{u}$. The independent components contribute to the motion as an offset around the mean. Our experiments showed that changing the mean often results in unnatural motion or violation of physical constraints such as lip intersection. However, modifying the mixing matrix and/or the independent components yields interesting editing operations.

Translate In Section 4 and Figure 1(a) it was made clear that a single independent component captures the difference between *Frustrated* and *Happy* motions. Moreover this component seems to vary between two extreme values as a function of emotion. A straightforward way of modifying emotion is then to estimate these extreme values and translate the time series responsible for emotion between them. With this technique we can change the emotion continuously between the two emotions present in the training set. Editing can be expressed as:

$$\boldsymbol{x} = E\{\boldsymbol{x}\} + \boldsymbol{PA}(\boldsymbol{u} + \alpha \boldsymbol{e}_E),$$

where α is a scalar that quantifies the amount of translation in the emotional component and $\boldsymbol{e}_E$ is the vector in the canonical basis of the ICA mixing matrix that corresponds to the emotional component.

Copy and Replace Another editing operation is to replace the emotional component of a motion with the emotional

component of a different motion without changing the content (speech related motion) of the original motion. To do this we replace the time series that corresponds to the emotional component u_1 in ICA space by the emotional component of a second motion u_2. This manipulation can then be written as follows:

$$\boldsymbol{x} = E\{\boldsymbol{x}\} + \boldsymbol{PA}(\boldsymbol{u}_1 + ((\boldsymbol{u}_2 - \boldsymbol{u}_1)^T \boldsymbol{e}_E)\boldsymbol{e}_E).$$

Copy and Add We can also add an emotional component was is not present in the original motion. Let's consider u_1 and u_2 the emotional components of two motions. In order to add the emotional component of motion 1 to motion 2 we perform the following operation:

$$\boldsymbol{x} = E\{\boldsymbol{x}\} + (\boldsymbol{PA})_1\boldsymbol{u}_1 + (\boldsymbol{PA})_2((\boldsymbol{u}_2^T \boldsymbol{e}_E^2)\boldsymbol{e}_E^2),$$

where $(\boldsymbol{PA})_1$ and $(\boldsymbol{PA})_2$ are the mixing matrices of the two motions. $\boldsymbol{e}_E^2$ is the vector in the canonical basis of the ICA mixing matrix $\boldsymbol{A}_2$ that corresponds to the emotional component.

Notice that all the editing operations we have described so far are applied to motions that are already in the training set used to estimate the ICA model. In order to edit a motion $\boldsymbol{x}$ that does not belong to the training set, we can project it to extract the independent components:

$$\boldsymbol{u} = (\boldsymbol{PA})^+(\boldsymbol{x} - E\{\boldsymbol{x}_{training}\}),$$

where $^+$ indicates the pseudo-inverse of a matrix and $\boldsymbol{x}_{training}$ the expectation of the motions in the training set. After projection, the motion can be edited in ICA space.

6. Results

Our results show that ICA can be used to decompose speech related facial motion into meaningful components. In this section we discuss the results of several experiments. Please see the supporting video for a clearer demonstration of our results or http://www.cs.ucla.edu/∼abingcao.

Motion capture and rendering We recorded facial motion using a Vicon8 optical motion capture system. We used 109 markers to sample the face geometry fairly densely. The sampling rate of the data is 120 frame/sec. To drive a 3D textured face mesh, the markers are mapped to corresponding mesh points, and the rest of the mesh is deformed using *Radial Basis Functions* [20].

ICA vs PCA In our experiments the principal components correlate the speech related mouth motion with intense emotion related eyebrow motion. In contrast, the independent components are able to separate mouth and eyebrow motion to a much more meaningful degree. The independent component that captures the mouth motion contains limited eyebrow motion. We believe that this is correct since part of the eyebrow motion is actually related to the content of the speech, for example when stressing a point. In contrast, intense eyebrow motion is clearly related to emotion and not to the content of the speech.

Editing The proposed method provides an intuitive decomposition of facial motion that allows us to edit the apparent emotion of visual speech. Figure 3 shows 3 rendered frames from an editing session. The *neutral* and *sad* independent components are mixed with different percentages. Figure 4 shows a emotion session that change the emotional content by translating between *neutral*, *sad* and *angry*.

7. Conclusion and future work

In this paper we propose an unsupervised learning technique based on *Independent Component Analysis*. With ICA we extract meaningful parameters from recorded facial motions. Our method provides a representation of the data that has a number of important features. First this representation has much more intuitive semantics than the original data. Each independent component can be associated with a clear meaning. These components can be edited separately. Second it is significantly more compact; 6 independent components instead of the original motion data that has 327 parameters (i.e. three euclidian coordinates for each of the 109 markers). But perhaps the most intriguing contribution of this paper is to show that facial motions should lend themselves so easily to a linear decomposition, despite the complexity of the associated control system (the brain) and of the mechanisms responsible for these motions.

We predict that ICA will find its way in the computer graphics toolbox alongside other statistical techniques such as PCA or clustering. For future work we plan to analyze more complex facial motions. In our current experiments we have restricted our training set to domain limited motions. Each motion has specific emotional content. This helped us to interpret the data. We would like to experiment with motions that have unconstrained emotions. It would be interesting to see if ICA could still extract emotional components from such data. Another direction we are investigating is to extend the set of editing operations on the independent components to a wider repertoire of signal processing techniques.

8. Acknowledgement

We like to thank Wen C. Tien for his help on this paper. This work was partly funded by the Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

We would also like to thank Intel Corp. and Microsoft Corp. for their generous support through equipment and software grants.

References

1. http://www.cis.hut.fi/projects/ica/fastica/.
2. J.N. Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality amd Social Psychology*, 39:2049–2059, 1979.
3. M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image otions. In *Proceedings, International Conference on Computer Vision*, pages 374–381, 1995.
4. M. Brand. Voice puppetry. In *Proceedings of ACM SIGGRAPH 1999*, pages 21–28. ACM Press/Addison-Wesley Publishing Co., 1999.
5. M. Brand and A. Hertzmann. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192. ACM Press/Addison-Wesley Publishing Co., 2000.
6. C. Bregler, M. Covell, and M. Slaney. Video rewrite: driving visual speech with audio. In *SIGGRAPH 97 Conference Proceedings*, pages 353–360. ACM SIGGRAPH, August 1997.
7. E. Chuang, H. Deshpande, and C. Bregler. Facial expression space learning. In *Proceedings of Pacific Graphics*, 2002.
8. J. Cohn, J Lien, A. Zlochower, and T. Kanade. Feature point tracking by optical flow discriminates subtle differences in facial expression. In *Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396–401, 1998.
9. I. Essa, S. Basu, T. Darell, and A. Pentland. Modeling, tracking and and interactive animation of faces and heads using input from video. *Computer Animation Conference*, pages 68–79, June 1996.
10. I. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. In *Proceedings of IEEE Computer Vision Pattern Recognition Conference*, 1994.
11. I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, July 1997.
12. T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of ACM SIGGRPAH 2002*, pages 388–398. ACM Press, 2002.
13. Z. Grahramani and M. I. Jordan. Factorial hidden markov models. *Machine Learning*, (29):245–275, 1997.
14. A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.
15. A. Hyvärinen. Survey on independent component analysis. In *Neural Computing Surveys*, pages 94–128, 1999.
16. Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *SIGGRAPH 95 Conference Proceedings*, pages 55–62. ACM SIGGRAPH, August 1995.
17. J. P. Lewis. Autmated lip-sunc: Background and techniques. *Journal of Visualization and Computer Animation*, 2:118–122, 1991.
18. J. Lien, J. Cohn, T. Kanade, and C.C. Li. Automatic facial expression recognition based on FACS action units. In *Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 390–395, 1998.
19. H. Mori and J. Hoshino. Independent component analysis of human motion. In *Proceedings of International Conference on Acoustics Speech and SIgnal Processing*, volume VI, pages 3564–3567, 2002.
20. T. H. Nelsons. *Literary Machiness*. Sausalito, CA: Mindful Press, 1993.
21. F. I. Parke. A model for human faces that allows speech synchronized animation. *Journal of Computers and Graphics*, 1(1):1–4, 1975.
22. K. Pullen and C. Bregler. Motion capture assisted animation: texturing and synthesis. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 501–508. ACM Press, 2002.
23. D. Sankoff and J. B. Kruskal. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. CSLI Publications, 1983.
24. J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.
25. M. Unuma, K. Anjyo, and R. Takeuchi. Fourier principles for emotion-based human figure animation. In *Proceedings of ACM SIGGRAPH 1995*, pages 91–96. ACM Press, 1995.
26. K. Waters. A muscle model for animating three-dimensional facial expression. In *SIGGRAPH 87 Conference Proceedings*, volume 21, pages 17–24. ACM SIGGRAPH, July 1987.

Figure 3: *Three snapshots of editing visual speech. Each row shows the same speech content but different amount of* neutral *and* sad *emotions.*

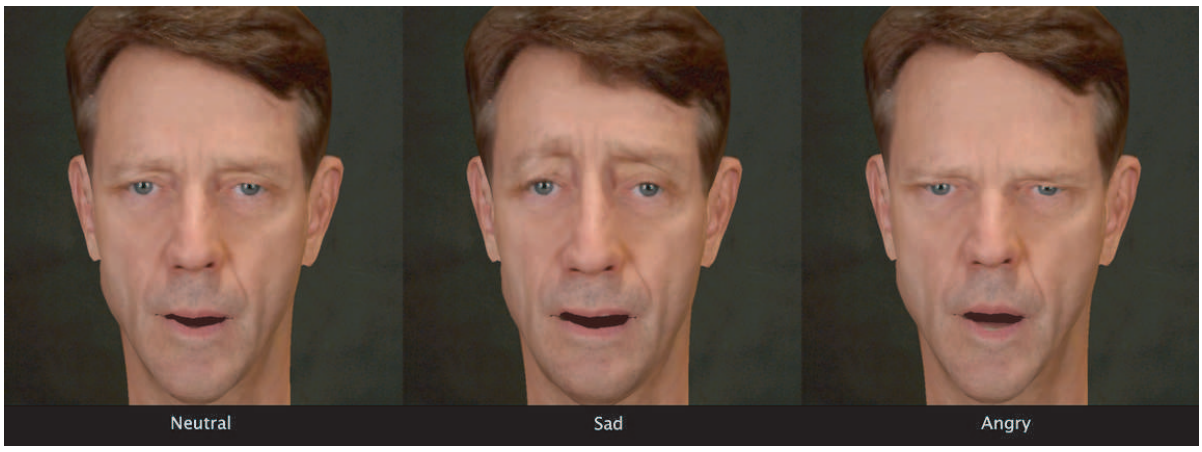

Figure 4: *Translating between three emotions*, neutral, sad *and* angry.