

# Facial Expression Space Learning

Erika S. Chuang, Hrishikesh Deshpande, Chris Bregler  
[echuang | hrishi | bregler@graphics.stanford.edu]  
<http://graphics.stanford.edu/~echuang/espace>  
Stanford University

## Abstract

*Animation of facial speech and expressions has experienced increased attention recently. Most current research focuses on techniques for capturing, synthesizing, and retargeting facial expressions. Little attention has been paid to the problem of controlling and modifying the expression itself. We present techniques that separate video data into expressive features and underlying content. This allows, for example, a sequence originally recorded with a happy expression to be modified so that the speaker appears to be speaking with an angry or neutral expression. Although the expression has been modified, the new sequences maintain the same visual speech content as the original sequence. The facial expression space that allows these transformations is learned with the aid of a factorization model.*

## 1 Introduction

Animation of facial speech and expressions has experienced increased attention recently. Most current research focuses on techniques for capturing, re-synthesizing, and re-targeting motion [3][5][16][25][26][31][35]. These methods record an actress/actor and then either create a new animation sequence in the style of the original, or adapt the motion to a new face model. Not much attention has been paid to the problem of controlling and modifying the facial expression itself. If the actor/actress was originally recorded in a happy mood, all output animations are generated with the same happy expressions. Ultimately, life-like and realistic facial animation needs to cover the entire range of human expression and mood change. Of course it is possible to collect all possible facial motions in every possible mood, and then apply existing techniques. Unfortunately, this requires a very large amount of data. Instead, we present techniques that derive the expressive style from video training data, and are able to induce different expressions onto existing data. For instance, a recording with happy expression can be transformed into the same spoken sequence, but with an angry expression.

A great deal of previous research involves the study of appropriate models of facial expression. Most of the effort has been with regard to tracking and recognition of facial expressions. These methods primarily focus on the static or short-term dynamics of facial expressions. By short-term dynamics we refer to the timing of a single action, such as a smile or a frown, where the facial configuration starts at a somewhat neutral position and undergo some changes to reach a final state. In contrast, very little work exists on the study of longer expressive motions.

In this work, we focus on the effect of various expression components, or emotion, during continuous speech. This is important because a realistically animated character will need to talk and convey emotions at the same time, instead of talking and then stopping intermittently just to make expressive faces. The additional challenge comes from the fact that as one tries to make a certain mouth shape necessary for speaking, the dynamics of the entire facial configuration change depending on the viseme. This is partially the reason that in natural speech, people do not maintain a particular expression with a constant degree, e.g. a happy person does not speak with a constant smile, instead a smile is apparent only at the appropriate points in a sequence. Modeling this variation is crucial to robust expression recognition as well as to synthesize natural looking facial animation.

To model this complicated interaction, we assume that at each time instant, the facial configuration is influenced by two underlying factors: a visual speech component and an expression component. We will call these factors the content and style of the facial configuration, following the terms used in the machine learning literature.

Many techniques exist that decipher data under influence of multiple factors [22][18][1][9][19]. We propose using the factorization techniques that have been studied extensively in the statistical learning, pattern recognition, and computer vision communities [14][15][29][30][33]. This technique formulates the content and style as a bilinear mapping. Bilinear models are two-factor models with the property that the outputs are linear in either factor when the other is held constant. Together, the two factors modulate each other's contributions multiplicatively, which allows rich

interactions between them. We believe that factorial models provide a good representation for coding variations in visual speech.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the factorization model and its application to facial expression learning. Section 4 provides the details of the experiment and results from analysis, synthesis tasks.

## 2 Related work

A great deal of research effort for facial expression analysis has been in the area of recognition. This work typically involves visual tracking of: facial movement such as contours, optical flow, or transient features such as furrows and wrinkles [12][13][21][32]. Facial expressions are learned and recognized in relation to a Facial Action Coding System (FACS) [11], a popular representation that codes visually distinguishable facial movements in small units. These action units describe qualitative measures such as “pulling lip corners”, or “frown deepening”. A facial expression is described as a combination of these units. Because of the lack of temporal elements in this analysis, many researchers have added a temporal dimension to their systems to improve the performance. This method is limited to analyzing single facial expressions such as a smile, or a frown. For animating facial speech, these descriptors, or action units are still insufficient. The MPEG4 standard extends this system to derive a set of Facial Animation Parameters (PAR) [24]. In addition to single expressions, facial speech is achieved by a set of viseme components. Systems that use this standard have achieved good compression for facial animation [10]. However, it is difficult to reliably estimate these physical facial parameters from video, and subtle details of facial expressions such as wrinkles and furrows are often lost in this representation. While these models are adequate for recognition, this work seeks a more expressive model suitable for synthesizing new animations.

In recent years, people have tried to bypass some of the problems with estimating physical facial parameters and chosen to animate faces directly using source data. Several methods for audio driven facial animation fall in this category. For example, Video Rewrite uses existing footage to automatically create new video of a person saying words that she did not say in the original footage [5]. It does so by re-ordering the existing video frames. Voice Puppetry introduces a method that predicts the facial control signal from the audio signal [3]. Full facial dynamics is learned from video data, and driven from audio via a probabilistic model. The focus of these methods is on the mapping between audio and visual speech signals and effects such as co-articulation to create natural speech. In this paper we address the issue of

learning mappings between the visual components of several expressions, rather than between audio and visual components of a single expression.

Goals similar to those of this work have been investigated in the domain of full body animation. Brand and Hertzman used a Stylistic Hidden Markov Model to model dance such that styles can be learned and applied to novel sequences [4]. Pullen and Bregler decompose motion into different frequency bands, and use the low frequency signal as the baseline for synthesizing new motion, and the high frequency signals to ‘texture’, or enhance the style and personality of new motions [27]. This work investigates separating facial rather than full body animation into content and style.

Finally, there are many methods in the autonomous agent community that generate animated virtual characters with emotions [6][20]. Most of the systems are rule based, where the rules are either results from behavioral studies, or derived from audio speech signals. In this work we seek a more expressive and detailed model than is typically available from rule based systems.

## 3 Facial expression model

This section describes the facial expression model, beginning with the representation for the facial configuration and followed by the description of the factorization model, also known as bilinear model. In the rest of the paper, the words factorization and bilinear are used interchangeably to refer to the same method.

### 3.1 Face representation

We define a facial configuration as the shape and texture of the face at a given instant of time. Shape is represented by a vector of feature locations on the face,  $x$ , and texture by the pixel intensity values of the image,  $g$ . Without reduction, this facial configuration vector may have thousands of elements since it includes each feature location as well as all pixel intensities. To minimize the dimensionality of the input vector, we represent the face images by a statistical model of the shape and color appearance similar to [7]. A set of  $n$  tracked feature points,  $f_1, \dots, f_n$ , on the face define the face shape, represented as the vector  $\bar{x}$ , where  $\bar{x} = [f_1^x, f_1^y, \dots, f_n^x, f_n^y]^T$ . We apply principal component analysis (PCA) to the data from all frames to obtain a representation with reduced dimensionality. Each example shape can now be approximated using:

$$\bar{x} = \bar{x}_0 + \mathbf{P}_s \bar{b}_s \quad (1)$$

where  $\bar{x}_0$  is the mean shape vector,  $\mathbf{P}_s$  is the set of orthogonal modes of variation derived from the PCA, and  $\bar{b}_s$  is a set of shape parameters. This new representation has a flexible dimensionality according to the number of principle components that are retained.

The dimensionality of the pixel intensity data can be reduced via a similar method. To build a statistical model of color appearance we warp the images to be aligned with the mean shape, and then sample the intensity values from the shape-normalized image. Similarly to the shape model, we apply PCA to the texture data,  $\bar{g}$

$$\bar{g} = \bar{g}_0 + \mathbf{P}_g \bar{b}_g \quad (2)$$

where  $\bar{g}_0$  is the mean gray-level vector,  $\mathbf{P}_g$  is a set of orthogonal modes of variation and  $\bar{b}_g$  is a set of color parameters.

The shape and appearance of the face can thus be summarized by the vectors  $\bar{b}_s$  and  $\bar{b}_g$ . The facial configuration vector at any given frame is defined as

$$\bar{y} = \begin{bmatrix} \bar{b}_s \\ \bar{b}_g \end{bmatrix} \quad (3)$$

In this work, the number of modes was chosen empirically. We used 32 modes of variation to represent the range of facial feature poses, and 16 modes of variation for each color channel.

### 3.2 Factorization model

We would like to study the structure of facial configurations, when bound together by either common expressions, or common visual speech components, in order to acquire general information that will be useful for analyzing novel observations. A factorization model represents the relationship between expression, visual speech content, and general information as matrix multiplication as follows.

Given  $\bar{y}^{sc}$  as the facial configuration vector for expression  $s$  and content  $c$ , then the  $k$ -th element,  $y_k^{sc}$ , can be described as,

$$y_k^{sc} = \bar{a}^s \cdot \mathbf{W}_k \cdot \bar{b}^c \quad (4)$$

One vector  $\bar{a}^s$  codes a specific style of expression, such as happy or angry, while the other vector  $\bar{b}^c$  codes the visual speech content, e.g. the current viseme. The

$k$ -th observed facial configuration vector,  $y_k^{sc}$ , with style  $s$  and content  $c$  is a weighted multiplication of the expression vector  $\bar{a}^s$ , the content vector  $\bar{b}^c$ , and a matrix  $\mathbf{W}_k$ . Note that  $\mathbf{W}_k$  is a weighting matrix that is constant across all content and expression vectors. This content and style independent matrix can be used to characterize the features common to all samples in the training set. For  $K$  dimensional vector  $\bar{y}^{sc}$ ,  $K$  of such matrices is required to describe the vector. In this case,  $K=48$  because of the dimension of facial representation we have used.

In some applications one of the factors is known in advance. For example, we may know that a new facial configuration has a particular style. In these cases a simplified model referred to as asymmetric bilinear can be used [29][30].

$$\bar{y}^{sc} = \tilde{\mathbf{W}} \cdot \bar{b} \quad (5)$$

$\tilde{\mathbf{W}}$  is either a content-specific or a expression-specific matrix, depending on which way the input is defined. If  $\tilde{\mathbf{W}}$  is the expression-specific matrix, the facial configuration is then the product of the this basis with a content vector. In this paper,  $\tilde{\mathbf{W}}$  is interpreted as the expression-specific matrix, although the other way is also equally valid.

The details of obtaining the weighting matrices  $\mathbf{W}_k$  and  $\tilde{\mathbf{W}}$  are described in section 4.2 .

## 4 Experiment

The model described above can be used to change the apparent expression of video sequences. Training data is acquired and analyzed to create the model. Given the model and footage with new content, a sequence can then be synthesized with any desired expression. In the experiment described here the training data consists of about 12 seconds of video recording from text in ‘‘Alice in Wonderland’’. The same textual content was performed by an actress in 3 expressions: neutral, happy, and angry. The remainder of this section describes: video tracking, model training, and analysis and synthesis of new sequences.

### 4.1 Video tracking

We need good tracking of the facial contours to train the appearance model used for facial representation. In order to capture the full expressiveness, the actress was allowed to move her head while speaking. Therefore, there was quite a bit of head motion during speech, making the tracking more challenging.

For tracking the face contour, we apply a technique called Space-Time tracking [34]. Based on the assumption that the non-rigid motion of an object can be factorized into a rigid motion and deformation described by a linear combination of basis shapes, a global low-rank constraint on the measurement-matrix can be established. This rank deficiency can be exploited to infer a basis of trajectories from the tracks of a few reliable feature points. The estimated basis allows us to transform the tracking problem for additional features into a search in the low dimensional trajectory subspace. An example of tracked image is shown in figure 1. In this case, reliable points include the three marker points on the actress’ cheek and chin, the tip of the hairline, and the outer corners of the eyebrows, where there are prominent



**Figure 1: Example of tracked facial contours.**

corner features that are correlated to the motion of the face contour.

Some facial contours, such as the lips, do not have reliable features with good correlation to the non-rigid facial motion. For these contours we use a model-based tracking technique related to EigenTracking [2]. From one frame to the next, the object of interest undergoes changes both in the affine motion and the coefficients of the eigenface model. Tracking the object is formulated as an optimization problem that searches for the best affine parameters and basis coefficients. The image database necessary for the training of the appearance model is obtained using the method of Covell and Bregler [8]. A handful of images from the sequence are labeled with contour points. Then, each labeled image is morphed with every other labeled image to enlarge the database. The eye and eyebrow contours were also tracked using this method.

## 4.2 Model training

In section 3.2, symmetric and asymmetric factorization models were described. We will first

describe the training process for the asymmetric bilinear model because it is simpler than the symmetric model.

Consider an example where the training data consists of neutral, happy, and angry expressions, and with content frames 1 through  $C$ , we first represent the facial data as facial configuration described in section 3.1. The input vectors are stacked in the vertical direction for the different expressions, and horizontally for the different content frames to form a single matrix  $\mathbf{Y}$ .

$$\mathbf{Y} = \begin{bmatrix} \bar{y}^{N1} & \dots & \bar{y}^{NC} \\ \bar{y}^{H1} & \dots & \bar{y}^{HC} \\ \bar{y}^{A1} & \dots & \bar{y}^{AC} \end{bmatrix} \quad (6)$$

This formulation requires common content vectors from different expressions to be in the same columns. However, people generally speak at different tempos for different facial expressions, thus correspondence cannot be directly established in the raw video frames. We solve for this correspondence by temporally aligning the audio speech signal to match the video frames, since visual speech components are likely to be correlated with the audio speech. The speech signal is processed using Relative Spectral Perceptual Linear Prediction (Rasta-PLP) [17]. Dynamic time warping, commonly used in speech recognition [28], is then applied to align all recordings of the same speech content, captured with different facial expressions. After warping the sequences, correspondence can be established between frames in different expressions. In our experiment, the recorded frames from the happy and angry sequences were temporally aligned to the neutral expression, so that all sequences have  $C$  frames.

For the asymmetric model, this input matrix can be factored using singular value decomposition (SVD),

$$\text{SVD}(\mathbf{Y}) = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^t \quad (7)$$

Then,

$$\mathbf{Y} \cong \mathbf{W} \cdot \mathbf{B} = \begin{bmatrix} \mathbf{W}^N \\ \mathbf{W}^H \\ \mathbf{W}^A \end{bmatrix} \cdot \mathbf{B} \quad (8)$$

$\mathbf{W}$  is the first  $J$  columns of  $\mathbf{U} \cdot \mathbf{S}$ , and  $\mathbf{B}$  is the first  $J$  rows of  $\mathbf{V}^t$ .  $\mathbf{W}^N$ ,  $\mathbf{W}^H$ , and  $\mathbf{W}^A$  are expression-specific submatrices that act as neutral, happy and angry expression basis respectively, as described in section 3.2. Note that each expression-specific basis is not an orthogonal basis by itself. Instead, the structure of the model ensures corresponding basis vectors across different expressions.

Although the asymmetric model is easier to train it assumes the expression of new data to be analyzed is known. The symmetric model allows more generality

since neither the expression nor content of new data needs to be known a priori.

For the symmetric model, it has been shown that an iterative fitting algorithm can learn  $\bar{a}^s$ ,  $\mathbf{W}_k$ , and  $\bar{b}^c$  for problems of this sort [23]. This algorithm, in essence, iteratively applies Singular Value Decomposition (SVD) to a simplified sub-problem that finds only  $\bar{a}^s$  or  $\bar{b}^c$  and repeatedly alternates between the two. It converges to a solution that has the minimum mean squared error relative to the training set across all samples.

For the facial configuration example, the training data can be written as

$$\mathbf{Y} = \begin{bmatrix} \bar{y}^{N1} & \dots & \bar{y}^{NC} \\ \bar{y}^{H1} & \dots & \bar{y}^{HC} \\ \bar{y}^{A1} & \dots & \bar{y}^{AC} \end{bmatrix} = [\mathbf{W}^{VT} \cdot \mathbf{A}]^{VT} \cdot \mathbf{B} \quad (9)$$

$\mathbf{A}$  is an expression matrix with each column representing a specific style, and  $\mathbf{B}$  is a matrix coding content specific column vectors. The matrix  $\mathbf{W}$  is a stack of  $K$  matrices  $\mathbf{W}_k$  defined in equation (4), and  $\mathbf{W}^{VT}$  is the vector transpose of  $\mathbf{W}$ . To fit the model, we first apply SVD to  $\mathbf{Y}$  as in the asymmetric case. The first  $J$  row of  $\mathbf{V}^t$  can be used as the initial estimate for  $\mathbf{B}$ . To solve for  $\mathbf{A}$ , we stack  $\mathbf{Y}$  in an alternate configuration known as the vector transpose resulting in the roles of content and style being reversed.

$$\mathbf{Y}^{VT} = \begin{bmatrix} \bar{y}^{N1} & \bar{y}^{H1} & \bar{y}^{A1} \\ \vdots & \vdots & \vdots \\ \bar{y}^{NC} & \bar{y}^{HC} & \bar{y}^{AC} \end{bmatrix} = [\mathbf{W} \cdot \mathbf{B}]^{VT} \cdot \mathbf{A} \quad (10)$$

After some matrix manipulation, we have  $[\mathbf{YB}]^{VT} = \mathbf{W}^{VT} \mathbf{A}$ . Another SVD can now be performed on  $[\mathbf{YB}]^{VT}$ , where  $\mathbf{B}$  was found during the previous estimation. The result of this SVD can be used to define the  $\mathbf{A}$  matrix as the first  $I$  rows of  $\mathbf{V}^t$ . Iterating on this procedure will

result in the convergence of values for both  $\mathbf{A}$  and  $\mathbf{B}$ . After convergence, the  $\mathbf{W}$  matrix can be found by

$$\mathbf{W} = \left[ [\mathbf{YB}^T]^{VT} \mathbf{A}^T \right]^{VT} \quad (11)$$

The choice of  $I$  and  $J$  determine the number of dimensions coded in the style and content vectors. In our experiment, we choose  $I=3$  to be equal to the number of expressions in the training set to allow maximum expressiveness. The dimensionality of  $J=10$  is chosen experimentally to allow reasonable speed to convergence, but still allows enough degree of freedom for expressiveness. In the next section, we will discuss the problems that might occur when  $J$  is not chosen properly.

### 4.3 Analysis, extrapolation and translation

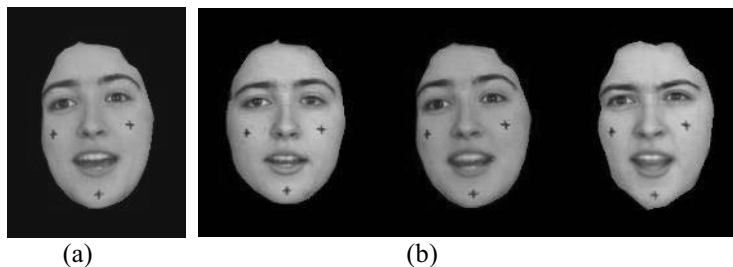
Given a new test image,  $\bar{y}_{test}$ , represented by the facial configuration shown in section 3.1, we would like to analyze it and perform some modification to the original data. For example, we would like to analyze it and obtain the content parameters  $\bar{b}_{test}$  and change the expression. If the expression style  $S$  is known, the content parameters are estimated via simple least square fit to the expression-specific parameters  $\mathbf{W}^s$  found from the training of the asymmetric model.

$$\bar{y}_{test}^S = \mathbf{W}^S \cdot \bar{b}_{test} \quad (12)$$

Extrapolating this test sequence to a new expressions,  $S_2$ , now simply requires multiplying the content,  $\bar{b}_{test}$ , with the style parameters for  $S_2$ .

$$\bar{y}_{test}^{S_2} = \mathbf{W}^{S_2} \cdot \bar{b}_{test} \quad (13)$$

To produce the output image, we simply take the



**Figure 2: example of analysis and extrapolation. (a) novel image in happy expression (b) in the middle is the reconstruction of the novel face in (a) using the bilinear model. The neutral face on the left and the angry face on the right are extrapolated from the model.**

portion of the facial configuration vector that depicts the shape and color parameters, and apply them to equation (1) and (2) to get the feature location and texture values. Since the texture is sampled from a mean shape, warping is applied to produce the final face image.

Figure 2 shows a single frame from an example result of analysis and extrapolation. The happy face in 2(a) is from an unseen image sequence. In figure 2(b), the middle face is the reconstruction of the novel face using the bilinear model. The left face is the extrapolated neutral face in the same content, and the right face is the extrapolated angry face. Figure 3 shows another example. In this case, a novel angry sequence was analyzed. The angry face in figure 3(a) is one frame from this sequence. In figure 3(b), the face on the right is the reconstruction of this angry frame, while the neutral and happy faces in the left are the extrapolated expressions.

Previously, we mentioned the importance of choosing the dimensionality of the content vector  $J$ . If it is too large, the extra dimensions essentially model noise in the data, which is not meaningful across expression bases. The result by applying  $\bar{b}_{test}$  to a different expression basis produced invalid images. If  $J$  is too small, there were not enough degrees of freedom in the model such that the results seem to be averaged out.

This method is robust enough that whole new sequences can be analyzed and new expressions extrapolated, despite the fact that frames are essentially treated independently. The model successfully captures the fact that facial expression is a dynamic event. A synthesized happy speaker smiles at only the appropriate points, rather than maintaining a constant facial shape bias.

Another task to which this expression model can be applied is to take a face image of unknown expression and unknown content, and extract both the expression and the content components. This is a more challenging task since the two factors interact with each other in a complex manner and the problem is very under-constrained. For

this task we use the symmetric formulation of the bilinear model from equation. Again, we chose the style dimensionality,  $I$ , to be 3, equaling the number of styles to allow maximum expressiveness. Therefore, only one iteration is required for the fitting algorithm (total of 3 SVDs during training). The content dimension  $J$  is set to 10 in this experiment.

To generalize a new test image  $\bar{y}_{test}$ , we adapt the model simultaneously for the new content  $c'$  and new expression  $s'$  iteratively. The mean content vector  $\bar{b}_0$  from the training set is used as the initial guess for the new content. We then iterate over the following 2 equations, where  $\mathbf{X}^{-1}$  is the pseudo-inverse of  $\mathbf{X}$ .

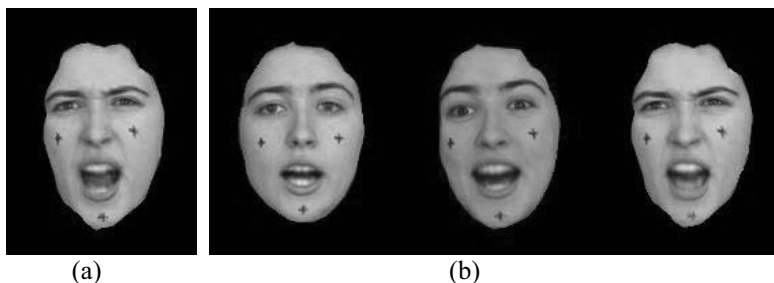
$$\bar{a}^{s'} = \left[ \left[ \mathbf{W} \bar{b}^{c'} \right]^{\text{T}} \right]^{-1} \bar{y}_{test} \quad (14)$$

$$\bar{b}^{c'} = \left[ \left[ \mathbf{W}^{\text{V}} \bar{a}^{s'} \right]^{\text{V}} \right]^{-1} \bar{y}_{test} \quad (15)$$

Figure 4 shows the results of this task. The top left corner is the novel image with unknown style and content. In this case, convergence occurs after about 50 iterations. To translate the new expression to any of the existing content or style classes from the training data, we simply replace  $\bar{a}^{s'}$  or  $\bar{b}^{c'}$  with the known content and style vectors. The top row shows the new style translated to known content. The left column shows the new content translated into known style.

#### 4.4 Adding constraints

Analyzing a sequence of images using the symmetric model is shown to be more difficult. Since the problem is under-constrained, if the new image is noisy, the alternating scheme in equation (14) and (15) sometimes does not converge to the correct local minimum. Additional constraints were applied in an ad-hoc fashion



**Figure 3 : example of analysis and extrapolation. (a) novel image in angry expression (b) on the right is the reconstruction of the novel face in (a) using the bilinear model. The neutral face on the left and the happy face in the middle are extrapolated from the model.**

to ensure the convergence to the correct local minimum. First, instead of solving for the content and expression parameters in the novel sequence frame by frame, we can solve for  $N$  frames at a time by assuming the expression vector  $\bar{a}^s$  does not change within these  $N$  frames. This is valid if we make the assumption that facial expression change at a slower rate than the frame rate. Second, by observing the fact that the first element of the content vector, derived from the training data, stays very constant over time, we can add that constraint in the algorithm to ensure a good initialization point. With the additional constraints, a correct local minimum can be achieved.

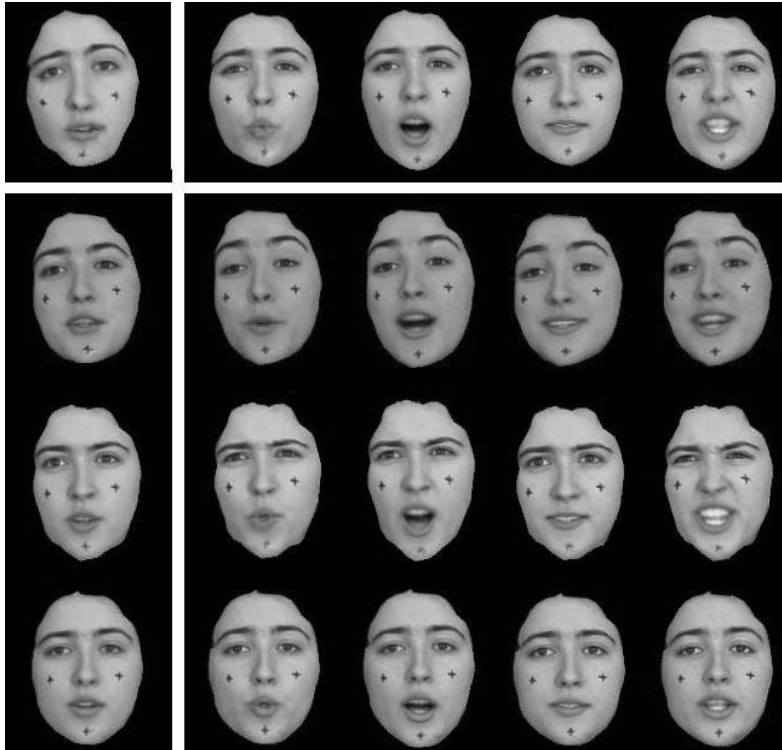
So far, the factorization model treats each frame individually, therefore the trajectory from frame to frame is sometimes not very smooth, resulting in jerky motion. A simple way to enforce temporal constraints to restrict frame-to-frame motion of the parameters to a valid dynamic range is to explicitly treat  $n$  frames as a bundle. For example, take  $n=3$ , the input configuration vector becomes

$$\bar{y}^{iS} = \begin{bmatrix} \bar{y}^{iS} \\ \bar{y}^{(i+1)S} \\ \bar{y}^{(i+2)S} \end{bmatrix} \quad (16)$$

where  $i$  is  $i$ -th column in the training matrix  $Y$  for expression  $S$  (equation (6)). When analyzing a new sequence, a sliding window of  $n$  frames is fitted to the model. To construct the output sequence, only the middle part of the configuration vector  $\bar{y}^{iS}$  is used to produce the image. The resulting animation from this modification appears to be a lot smoother.

## 5 Conclusion and future work

We show that the space of facial expressions can be modeled with a bilinear model. Two formulations of bilinear models have been fit to facial expression data. The asymmetric model derives style specific parameters, and is used to analyze a novel video sequence of known expression in order to recover the content. Together with the style-specific parameters for other expressions, new sequences of different expressions can be extrapolated.



**Figure 4: example of translation. The top left corner is the novel image with unknown style and content. The top row shows the new style translated to known content. The left column shows the new content translated into known style.**

The symmetric bilinear model derives a set of parameters independent of content and style. A novel image of unknown content and expression style is adapted to the model. Existing expressions with this new content, and existing content with this new expression are derived from this generalization.

There are many additional factors that contribute to expressive speech. For example, the tempo of speech is an important indicator of mood. In this work the timing of all sequences has been normalized. In the future we plan to investigate methods of including expression specific timing information in synthesized sequences. Additionally, patterns of expression and speech are correlated with respect to time. Finally, this model contains a number of arbitrary choices for dimensionality, both for reducing shape and texture content, as well as encoding content and style. In this work, these values were merely chosen in an ad hoc fashion. A more careful study of their effect may result an improved model.

## 6 Reference

- [1] Bell, A., Sejnowski, T. "An information-maximization approach to blind separation and blind deconvolution", *Neural Computation*, 1995.
- [2] Black, M. J. and Jepson, A., "EigenTracking: Robust matching and tracking of articulated objects using a view-based representation", *Proc. Fourth European Conf. on Computer Vision, (ECCV) 1996*.
- [3] M. Brand, "Voice Puppetry", *SIGGRAPH 1999*.
- [4] M. Brand, A. Hertzmann, "Style Machines", *SIGGRAPH 2000*.
- [5] C. Bregler, M. Covell, and M. Slaney, "*Video Rewrite: Driving Visual Speech with Audio*", *SIGGRAPH 97*.
- [6] J. Cassell, C. Pelachaud, N. I. Badler, M. Steedman, B. Achorn, T. Beckett, B. Douville, S. Prevost, M. Stone, "Animated conversation: rule-based generation of facial display, gesture and spoken intonation for multiple conversational agents" *SIGGRAPH 1994*.
- [7] T.F.Cootes, G.J. Edwards and C.J.Taylor. "Active Appearance Models", in *Proc. European Conference on Computer Vision 1998* (H.Burkhardt & B. Neumann Ed.s). Vol. 2, pp. 484-498, Springer, 1998.
- [8] M. Covell, C. Bregler. "Eigen-Points." *Proc. IEEE Int. Conf. on Image Processing*, 1996.
- [9] Dayan, P. Hinton, G. Neal, R. Zemel, R. "The Helmholtz machine", *Neural Computation*, 1995.
- [10] P. Eisert, T. Wiegand, and B. Girod, "Model-Aided Coding: A New Approach to Incorporate Facial Animation into Motion-Compensated Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 3, pp. 344-358, April 2000.
- [11] Ekman, P. & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, Calif.: Consulting Psychologists Press.
- [12] Essa, I. and S. Basu. "Modeling, Tracking and Interactive Animation of Facial Expressions and Head Movements using Input from Video", *Appears, Proceedings of Computer Animation 1996*
- [13] Essa, I., and A. Pentland. "Coding, Analysis, Interpretation and Recognition of Facial Expressions.", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 19 (7), IEEE Computer Society Press, July, 1997.
- [14] Z. Ghahramani. "Factorial learning and the EM algorithm." *Advances in neural information processing systems 7*, pp 617-624, MIT Press, 1995.
- [15] Z. Ghahramani, M.I. Jordan. "Factorial Hidden Markov Models." *Machine Learning*, 29, 245-275, 1997.
- [16] B. Guenter, C. Grimm, D. Wood, "Making faces", *SIGGRAPH 1999*.
- [17] Hermansky, H., Morgan, N., Bayya, A, and Kohn, P. "Rasta-PLP Speech Analysis", *ICSI Technical Report TR-91-069*, Berkeley, California.
- [18] Hinton, G. E., Zemel, R, "Autoencoders, minimum description length, and Helmholtz free energy", *Advances in neural information processing systems*, 6, 1994. Morgan Kauffmann.
- [19] Hinton G. and Ghahramani, Z. "Generative models for discovering sparse distributed representations". *Phil. Trans. Royal Soc. B*, 352, 1997.
- [20] Yan Li, Feng Yu, Ying-Qing Xu, Eric Chang, Heung-Yeung Shum, "Speech-Driven eCartoon Animation with Emotions", *ACM Multimedia*, 2001.
- [21] Lien, J.J., Kanade, T.K., Cohn, J.F., & Li, C.C. "A Multi-Method Approach for Discriminating Between Similar Facial Expressions, Including Expression Intensity Estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR98)*.
- [22] Marida, K. Kent, J., Bibby, J, "Multivariate Analysis", London: Academic Press, 1979.
- [23] D.H. Marimont, B.A. Wandell. "Linear models of surface and illuminant spectra." *J. Optical Society of America A*, 9(11):1905-1913, 1992.
- [24] J Ostermann, "Animation of Synthetic Faces in MPEG-4", *Computer Animation*, pp.49-51, Philadelphia, PA, June8-10, 1998.
- [25] Parke and Walters, "Computer Facial Animation", A.K. Peters, 1996.
- [26] Frederic H. Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, David Salesin: *Synthesizing Realistic Facial Expressions from Photographs*. *SIGGRAPH 1998*:
- [27] K. Pullen, C. Bregler, "Motion Capture Assisted Animation: Texturing and Synthesis", *SIGGRAPH 2002*.

- [28] Rabiner, Lawrence R. and B. H. Juang, "Fundamentals of speech recognition", Prentice Hall, 1993.
- [29] J. B. Tenenbaum , W. T. Freeman, "Separating Style and Content," in *Advances in Neural Information Processing Systems 9*, Morgan Kaufmann, 1997.
- [30] J. B. Tenenbaum, W.T. Freeman, "Separating style and content with bilinear models", *Neural Computation 12* (6), 1247-1283, 2000.
- [31] D. Terzopoulos, K. Walters, "Analysis and synthesis of facial image sequences using physical and anatomical models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1993.
- [32] Yingli Tian, T. Kanade and J. F. Cohn , " Recognizing Action Units for Facial Expression Analysis ", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, February, 2001.
- [33] C. Tomasi, T. Kanade. "Shape and motion from image streams under orthography: a factorization method." *Int. Journal of Computer Vision*, 9(2):137-154, 1992.
- [34] L. Torresani and C. Bregler, "Space-Time Tracking", *European Conference on Computer Vision, ECCV 2002*.
- [35] Lance Williams, "Performance-Driven Facial Animation", *In Computer Graphics (Proc. SIGGRAPH)*, pp. 235-242, 1990.