EG UK Theory and Practice of Computer Graphics (2011)
Hamish Carr, Ian Grimstead (Editors)

# Dynamic Video Face Transformation Using Multilinear and Autoregressive Models

Bernard Tiddeman[† 1], David Hunter[‡ 2] and David Perrett[§ 2]

[1] Department of Computer Science, Aberystwyth University, Wales, UK
[2] School of Psychology, University of St Andrews, Scotland, UK

**Abstract**

*In this paper we present a prototype system for altering perceived attributes of faces in video sequences, such as the apparent age, sex or emotional state. The system uses multilinear models to decompose the parameters coding for each frame into separate pose and identity parameters. The multilinear model is learnt automatically from the training video data. Statistical models of group identity are then used to alter the identity parameters from one group to another (e.g. from male to female). An autoregressive model is learnt from the pose parameters, and this is applied to alter the dynamics. We have tested our system on a small dataset (for altering apparent gender) with encouraging preliminary results.*

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation

## 1. Introduction

Altering the percieved attributes of faces in video sequences (such as age, sex or emotional state) has applications in psychology research and entertainment. In this paper we suggest that the alterations should take into account 3 types of change:

- Changes to the identity in a neutral or average pose
- Changes that also account for each particular expression, but treat each frame independently
- Changes that also alter the dynamics, by taking into account differences between frames.

Previous work has generally focussed on alterations to the appearance of each frame independently [TP02] [BBPV03] [VBPP05]. Multilinear models can be used to decompose the appearance in each frame into paramters that encode different attributes (such as pose and identity) [VT02] [MVV06] [VBPP05]. Multilinear models based on static frames or posed expressions do not take into account differences in the dynamics between different groups of subjects. Dynamic alteration of facial video clips has recently been addressed by [AKC*11] using a dynamic time warping (DTW) approach. The DTW approach is suitable for matching different sequences of (approximately) the same sequence of expressions, e.g. saying a fixed sentence, but requires further work to enable general purpose alterations to dynam-

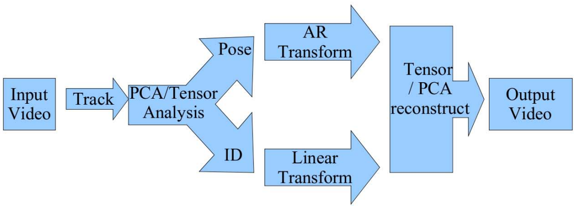

**Figure 1:** *A diagram of the transformation process. An input video is tracked, each frame is converted to PCA and then tensor components. The components are split into those that code for "pose" (includes out-of-plane rotations and expressions) and those that code for identity. The identity components undergo a linear transform based on group averages, the pose components are altered via the auto-regression approach. The altered model components are then used to reconstruct the output video.*

[†] Welsh Research Institute of Visual Computing
[‡] Unilever Research
[§] British Academy Wolfson Professorship

DOI: 10.2312/LocalChapterEvents/TPCG/TPCG11/049-052

ics. Here we attempt to devise a system capable of altering clips with an unknown sequence of expressions. Wampler *et al.* [WSZP07] also used a multilinear model built from short dynamic 3D clips (*animes*), where the focus was on animation of a previously unseen static mesh driven by annotated audio and emotional signals, rather than on alteration of existing clips as described here. In this paper we combine a multilinear decomposition with an autoregressive model of the facial dynamics in order to more accurately reflect the dynamic changes between groups.

To build the multilinear model a set of matching facial description vectors are required along each dimension (in this work just pose and identity). Finding a set of matching facial poses performed by different subjects from video sequences manually is extremely challenging. Instead we employ a simple scheme for building inputs to the multilinear models automatically. Once the multilinear model is built it can be used to decompose an input frame into parameters that code for different aspects such as pose and identity. To transform we treat the identity and pose parameters differently. A simple linear model is used to alter the identity parameters from one group to another. For the pose parameters two auto-regressive models are used to predict the next frames for the input and output sequences. The difference between the prediction and the actual value for the input sequence is used to adjust the predicted values for the output sequence. An overview of the process is shown in Figure 1.

The main contributions of this paper are therefore (1) to extend identity (e.g. age, gender etc) facial transforms to video taking account of dynamics, (2) automatic matching of video frames for multilinear model construction and (3) alteration of dynamics using an auto-regressive approach.

## 2. Method

### 2.1. Tracking

In the work described here the face is tracked using active appearance models (AAMs) [CET01]. AAMs uses Principal Component Analysis (PCA) of shape (as defined by landmarks on the training images) to automatically learn the main axes of variation in the training set. PCA is also applied to the image data by first warping all images into the average shape. The shape and colour principal components can be used to synthesise facial images by first constructing the shape and colour parts separately, and then warping the image part from the mean shape to the target shape. An additional PCA can be performed on the concatenated (and appropriately weighted) components, to exploit correlations between shape and appearance to create a more compact and class specific model.

Several algorithms have been developed to fit an AAM to an image e.g. [CET01] [MB03] [GMB05] [DRG*06]. AAMs typically only track sufficiently well when the training data closely matches the target data, for example when it is built from a sample of frames from the target video clip. In this work we use an NCC based AAM tracker [TC07], in a user interface that allows easy correction, rebuilding of the AAM and retracking to achieve good tracking results.

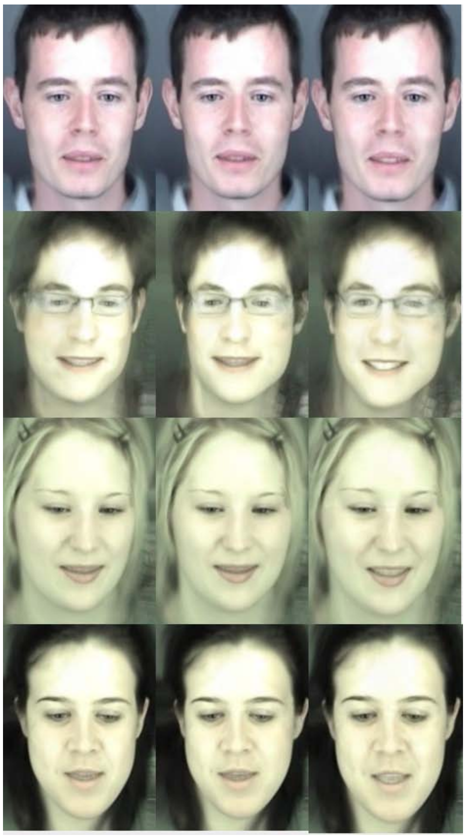

**Figure 2:** *Examples of the automatically generated matching poses for different subjects. Each row represents the model components for one subject, each column represents a different component. The components used to generate these images (with additional subjects and more components) are used to build the tensor model.*

### 2.2. Building Multilinear Models

After tracking the sequences we perform a multilinear analysis in PCA space, using a global PCA model to decompose each subject's image and template into a parameter vector. Multilinear models require a collection of data vectors arranged in an multi-dimensional grid (a tensor) with meaningful axes such as identity, pose, expression, lighting etc. These can be decomposed using multilinear PCA (MPCA) to give a more compact model. The decomposition requires unfolding of the tensor along one axis to give a 2D tensor (a matrix), to which PCA can be applied. The data is then folded back into the tensor and the process repeated along

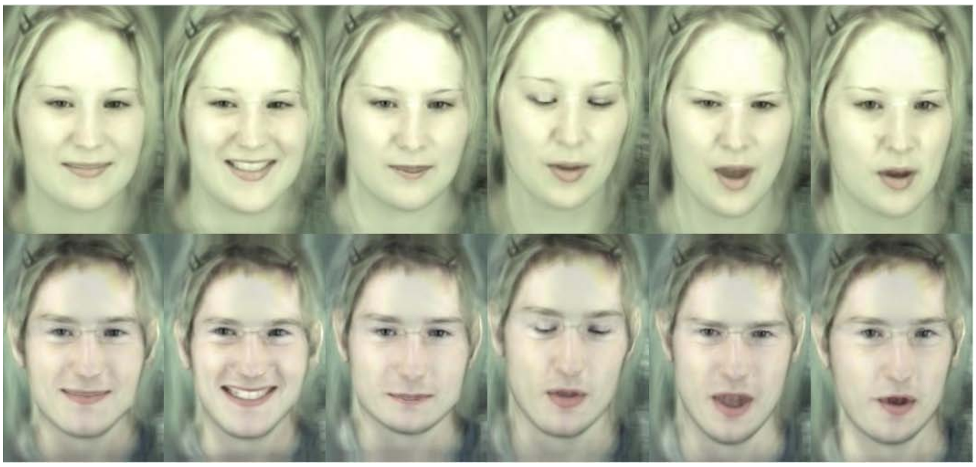

**Figure 3:** *Example results of transforming video frames. Top row shows the reconstructed images from the original tensor parameters, the bottom row shows the reconstructed images from the transformed tensor parameters, based on a female to male transformation.*

another axis. The process is repeated along all axes except the data axis. As with PCA the mean image should be subtracted before applying the MPCA. In this work we subtract the "fibre" mean as described in [THY07].

Before the MPCA can be applied a matching data vector from each subject in each condition is required. Matching frames from video by hand is not only difficult, it may in some instances be impossible (due to a lack of appropriate data), but a frame synthesised from a linear combination of the available frames might suffice. In addition it is not always clear what poses should be chosen for the model, and might be subject to a degree of perceptual bias leading to poor matches between subjects. For example small head rotations might alter the projected shape of the face significantly, but have a small impact perceptually. In order to automate the process we use a two step approach; first we identify the main axes of variation across the set and then we find the best match to each of these axes using data from each subject separately. To identify the main axes of variation we find the difference vector from each subject's own mean across the training set, i.e. we find the vectors:

$$\delta_{ij} = \mathbf{x}_{ij} - \frac{1}{N_j}\sum_{i=0}^{N_j} \mathbf{x}_{ij} \qquad (1)$$

and perform PCA analysis on the vectors $\delta_{ij}$. This gives us the target set of pose vectors to model with the tensor. Next we find the nearest projection of each global principal component into each subject's PCA space i.e. if $\mathbf{A}_j$ is the matrix with columns $\delta_{ij}$ from subject $j$ we solve:

$$\mathbf{A}_j \mathbf{w}_{jk} = \mathbf{p}_k \qquad (2)$$

where $\mathbf{w}_{jk}$ is the vector of weights to put on the residual vectors $\delta_{ij}$ for subject $j$ to best approximate the $k^{th}$ global principal component $\mathbf{p}_k$. This is solved in the least squares sense using:

$$\mathbf{w}_{jk} = (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\mathbf{p}_k \qquad (3)$$

This gives the best (in the linear least squares sense) approximation to the global combined appearance vector possible by linear combinations of the specific individual's pose vectors. The length of the vector for each subject is chosen to be the standard deviation of projections of that subject's residual vectors onto the estimated direction i.e.

$$l_{jk}^2 = \sum_{i=0}^{N_j} (\hat{\mathbf{p}}_{jk} \cdot \delta_{ij})^2 \qquad (4)$$

where $l_{jk}$ is the length of the $k^{th}$ ouput synthetic vector for subject $j$ and $\mathbf{p}_{jk} = \mathbf{A}_j\mathbf{w}_{jk}$ gives the direction of the vector. These vectors are used as input to the multilinear training stage. Examples of the automatically matched examples are given in Figure 2.

### 2.3. Altering the identity

In order to affect alteration to specific percieved attributes we use a simple Gaussian model of the group identity parameters. By construction the identity parameters across the training videos are approximately constant. We take the identity parameters for each training subject and build a Gaussian model representing each group. In this work we experiment only with male and female as the two groups, but the methods could in principal be extended to any groupings, such as by age or emotional state. In this work we simply add the differences between the means to a subject's identity parameters to transform the identity parameters from one group to another. In the future we will experiment with more complex models that attempt to map the differences in the shape of the distributions (by rotation and scaling along the axes).

### 2.4. Auto-regressive modelling for dynamics

The final ingredient in our proposed system is to learn and vary properties relating to the dynamics. It is possible to learn the dynamics for individual specific sequences (for example by matching via dynamic time warping) but it is not clear how this can be extended out of set to novel sequences. A sufficiently large codebook of short sequences may allow matching to most longer sequences, but instead in this work we learn a generic model via a simple auto-regression of the previous N-frames. A linear model of the expected next frame is learnt from the previous N-frames for each group of subjects. The model is learnt only on the "pose" parameters as we do not expect the identity parameters to vary (much) across a video sequence.

To apply this model to alter a specific video clip we predict the next frame's parameters for both the input sequence and for the output sequence (using the previous N synthesised frames). The appropriate group AR model is used to form the prediction in each case. The actual next frame for the input sequence is known, and we assume that the next output frame differs from the prediction by the same vector amount. This gives us a simple linear model to alter the pose parameters, but does have some limitations. The main problem is that the length of sub-sequences (i.e. between key frames) will remain constant. Temporal resampling (based on the average group velocities) can be integrated into the predictive transform process to improve the model.

## 3. Preliminary Results

Video clips were captured of 6 individuals (3 male and 3 female) saying a "standard" short sentence, designed to illicit the major visemes in English. Each subject was asked to say the sentence in a number of emotional states, although the quality of the acting was insufficiently consistent to be used as the basis of transformations, hence we concentrate on altering perceived gender. The subjects were asked to face the camera, but some residual head rotations are included in the data.

Preliminary experiments have been very basic, and designed only to test that the system is operating correctly. These have involved altering both within set and out-of-set video clips. A typical result is shown in Figure 3.

## 4. Conclusions and Future Work

In this paper we have presented a prototype system for dynamic facial transformations based on tensor decomposition and an auto-regressive approach to altering dynamics. Preliminary results produced by the system have been presented. Future work will include collecting a much larger and higher quality training set and conducting a perceptual experiment to validate the technique. Initial indications are that the effects of the dynamic alterations are small but they may nevertheless improve the perceived results.

## References

[AKC*11] AUBREY A., KAJIC V., CINGOVSKA I., ROSIN P., MARSHALL D.: Mapping and manipulating facial dynamics. In *Int. Conf. on Automatic Face and Gesture Recognition* (2011). 1

[BBPV03] BLANZ V., BASSO C., POGGIO T., VETTER T.: Reanimating faces in images and video. *Computer Graphics Forum 22*, 3 (September 2003), 641– 650. 1

[CET01] COOTES T. F., EDWARDS G. J., TAYLOR C. J.: Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 6 (2001), 681–685. 2

[DRG*06] DONNER R., REITER M., GEORG L., PELSOCHEK P., HORST B.: Fast active appearance model search using canonical correlation analysis. *IEEE TPAMI 28*, 10 (October 2006), 1690–1694. 2

[GMB05] GROSS R., MATTHEWS I., BAKER S.: Generic vs. person specific active appearance models. *Image and Vision Computing 23*, 11 (2005), 1080–1093. 2

[MB03] MATTHEWS I., BAKER S.: Active appearance models revisited. *International Journal of Computer Vision 60* (2003), 135–164. 2

[MVV06] MACÊDO I., VITAL E., VELHO B. L.: Expression transfer between photographs through multilinear aam's. In *In Brazilian Symposium on Computer Graphics and Image Processing* (2006), pp. 239–246. 1

[TC07] TIDDEMAN B., CHEN J.: Correlated active appearance models. In *IEEE Conference on Signal and Image Technologies in Internet Systems* (2007). 2

[THY07] TIDDEMAN B., HUNTER D., YU M.: Fibre centred tensor faces. In *British Machine Vision Conference* (2007), pp. 449–458. 3

[TP02] TIDDEMAN B., PERRETT D.: Transformation of dynamic facial image sequences using static 2d prototypes. *The Visual Computer 18*, 4 (June 2002), 218–225. 1

[VBPP05] VLASIC D., BRAND M., PFISTER H., POPOVIC J.: Face transfer with multilinear models. *SIGGRAPH 24*, 3 (July 2005), 426–433. 1

[VT02] VASILESCU M. A. O., TERZOPOULOS D.: Multilinear analysis of image ensembles: Tensorfaces. In *Proc. of the European Conf. on Computer Vision (ECCV 2002)* (May 2002), pp. 447–460. 1

[WSZP07] WAMPLER K., SASAKI D., ZHANG L., POPOVIÄĞ Z.: Dynamic, expressive speech animation from a single mesh. In *ACM Symposium on Computer Animation (SCA)* (2007), Metaxas D., Popovic J., (Eds.), pp. 53–62. 2