Linear Facial Expression Transfer With Active Appearance Models

Miles de la Hunty[1], Akshay Asthana[1] and Roland Goecke[2,1]
[1]*School of Engineering, CECS and RSISE, Australian National University*
[2]*Vision and Sensing & HCC Lab / NCBS, Faculty of ISE, University of Canberra*
miles.dlh@gmail.com *aasthana@rsise.anu.edu.au* *roland.goecke@ieee.org*

Abstract

The issue of transferring facial expressions from one person's face to another's has been an area of interest for the movie industry and the computer graphics community for quite some time. In recent years, with the proliferation of online image and video collections and web applications, such as Google Street View, the question of preserving privacy through face de-identification has gained interest in the computer vision community. In this paper, we focus on the problem of real-time dynamic facial expression transfer using an Active Appearance Model framework. We provide a theoretical foundation for a generalisation of two well-known expression transfer methods and demonstrate the improved visual quality of the proposed linear extrapolation transfer method on examples of face swapping and expression transfer using the AVOZES data corpus. Realistic talking faces can be generated in real-time at low computational cost.

1. Introduction and Related Work

Motion capture and the transfer of facial expressions from an actor to a CGI-generated movie character have long been a focus of much research in the movie industry and computer graphics community. The quest is to copy the facial movements of the source as truly as possible, while presenting them in a plausibly looking way on the animated character. While early approaches required specially made up faces or artificial landmarks to facilitate the required face tracking and model alignment, technological advances in the last decade have enabled markerless solutions. Beyond the film making industry, such technology is also of interest for the purpose of preserving privacy through face de-identification in online image and video collections and web applications, such as Google Street View.

In this paper, the problem of facial expression transfer ('cloning') by linear extrapolation is investigated, within the *Active Appearance Model* framework. The problem can be stated as follows: Given AAMs of two subjects, how can we convincingly map expressions observed in one face onto the model of the other? We give a theoretical justification for a linear method not requiring statistical training, and thus suitable for transfer in real-time.

In recent years, statistical approaches, such as the Active Appearance Model (AAM) [4] and 3D Morphable Model (3DMM) [3], have been widely and successfully used for building non-rigid deformable models. Their power lies in the combination of a compact parametric representation and an efficient alignment method. A number of AAM alignment methods have been proposed in recent years. Without loss of generality, we adopt the discriminative-iterative approach of [6] for the face tracking.

Our work is primarily inspired by the work of Theobald *et al.* [7], Blanz *et al.* [2], and Bitouk *et al.* [1]. [7] proposes a simple mapping of parameters between AAMs for two or more people without requiring high-level semantic information about the facial expressions, which is capable of real-time expression transfer. Here, the relationship between the shape vectors of two AAMs is computed and used to determine how the parameters of one model are to be mapped to the parameters of the other model.

[2] presents a 3D model approach, which estimates 3D shape and texture along with all relevant scene parameters (e.g. pose, lighting) from single images. It is based on the 3DMM [3] that gives highly accurate face models, but is computationally very expensive. The advantage of the 3D model approach is the ability to handle different poses, in particular non-frontal ones, and lighting directions well.

[1] proposes a system for automatic face replacement using a large library of face images from the internet. Candidate face images that are similar to the input face in appearance and pose are selected from the face li-

brary, before the pose, lighting, and colour of the candidate face images are adjusted to match the source image. No 3D model is required.

2. Notation

Here, an 'appearance vector' describes the geometry (shape) and texture of a model instance, encoded for example as $(\mathbf{x}_0, \mathbf{y}_0, \mathbf{x}_1, \mathbf{y}_1, \ldots, \mathbf{r}_0, \mathbf{g}_0, \mathbf{b}_0, \mathbf{r}_1, \mathbf{g}_1, \mathbf{b}_1, \ldots)$ for a 2D RGB model, where each pair $\mathbf{x}_i, \mathbf{y}_i$ describes the location of a mesh vertex and each triple $\mathbf{r}_i, \mathbf{g}_i, \mathbf{b}_i$ describes the colour of a pixel in the untransformed texture layer.

For each AAM p, we denote $\overline{\mathbf{y}}_p$ to be the mean appearance vector of p, $\mathbf{Y}_p$ to be the basis of appearance variation, and $\mathbf{Q}_p$, $\overline{\mathbf{Q}}_p$ to be the projector and complementary projector along $\mathbf{Y}_p$[1], respectively.

3. Proposed Approach

3.1 Linear Expression Transfer

We first review two simple techniques by which expression transfer may be achieved. First, we may simply add the variation observed in the source frame to the target mean, with no adjustment whatsoever. If the observed expression vector is given as $\overline{\mathbf{y}}_s + \widehat{\mathbf{y}}_s$, then our synthesised vector will be $\overline{\mathbf{y}}_t + \widehat{\mathbf{y}}_s$.

In addition to being straightforward, it ensures fidelity to the observed expression: if the source subject raises an eyebrow, so too must the target, because exactly the same variation is applied in the affine warp. However, textural or geometric features of the source model that would be out-of-place in the target are also transferred, leading to occasional visual oddities. For example, if during speech, the teeth of the source model are more visible than in the target, we may end up with dimmed 'ghost teeth' appearing in the synthesised face.

Alternatively, we may restrict the synthesised expression to the existing expression space $\mathbf{Y}_t$ of the target model. Subject to this restriction, we may reproduce the expression with *minimal mean square error* between the appearance vectors by taking a simple linear projection of the observed expression into the target subspace: $\widehat{\mathbf{y}}_t = \mathbf{Q}_t \widehat{\mathbf{y}}_s$. Such an approach is described in [7], and a short proof of the MSE property is given in the appendix. In enforcing the existing expression space, this method prioritises consistency with our observations of the target model above fidelity to the observed expression variation. Any information present in the data not fully captured by the target model is simply discarded by the projection, thus tending to produce qualitatively 'conservative' reproductions. In the extreme case in which $\widehat{\mathbf{y}}_s$ is independent to *every* mode of variation in the target model, the vector of dot products $\mathbf{Y}_t^T \widehat{\mathbf{y}}_s$ is zero, and $\mathbf{Q}_t \widehat{\mathbf{y}}_s = \mathbf{Y}_t \mathbf{Y}_t^T \widehat{\mathbf{y}}_s = \mathbf{Y}_t \mathbf{0} = \mathbf{0}$ - all information is lost!

Generally though, we are interested in a reproduction, which is both *reasonably* plausible on the target model *and* maintains *most* of the information of the original expression. Thus, a compromise between the two approaches is necessary. The following parameterised cost function is sensitive to both these effects:

$$C(\mathbf{y}_t) = ||\mathbf{y}_t - \mathbf{y}_s||^2 + \varphi \cdot ||\overline{\mathbf{Q}}_t \mathbf{y}_t||^2 \quad . \qquad (1)$$

The first term measures the squared Euclidean distance between the variation from the mean of the original and synthesised expressions, offering a simple measure of *similarity* between the expressions. The second term uses the complementary projector $\overline{\mathbf{Q}}_t$ to extract the component of the synthesised expression variation $\mathbf{y}_t$ orthogonal to ('outside of') the target model's expression space, returning its square magnitude. The real parameter $\varphi > 0$ reflects the sensitivity to this effect.

Theorem 1 *Subject to the cost function defined in Eq. 1, the optimal replicating expression is given by:*[2]

$$\mathbf{y}_t^* = \left(\mathbf{Q}_s + \alpha \overline{\mathbf{Q}}_t\right) \mathbf{y}_s$$

where $\alpha = 1/(\varphi + 1) \in (0, 1]$.

A simple proof of this theorem is given in the appendix. Here, α determines the proportion of variation in the observed expression unavailable in the target model that we allow into the solution. Choosing $\varphi = 0$ and caring only about the Euclidean distance, gives the 'direct transfer' method with $\alpha = 1$ and $\mathbf{y}_t = \mathbf{y}_s$. As $\varphi \to \infty$, keeping within the expression space of the target model approaches a hard constraint; $\alpha \to 0$ and $\mathbf{y}_t^* \to \mathbf{Q}_t \mathbf{y}_s$, giving the simple 'projection' method. Between the extremes, we obtain the range of weighted averages of the two. Empirically, we found $\alpha = 0.6$ to provide a good balance.

[1] The projector $\mathbf{Q}_p$ is the matrix that preserves the range of $\mathbf{Y}_p$ ($\mathbf{Q}_p \mathbf{Y}_p = \mathbf{Y}_p$) and zeroes all other vectors. $\overline{\mathbf{Q}}_p = \mathbf{I} - \mathbf{Q}_p$ zeroes the range of $\mathbf{Y}_p$ and preserves everything else. For an introduction to matrix projectors, see for example [8], pp55-61. Note that as $\mathbf{Y}_p$ arises from PCA and hence is orthogonal, $\mathbf{Q}_p = \mathbf{Y}_p \mathbf{Y}_p^T$.

[2] As an implementation note, it is generally infeasible to explicitly form the matrices $\mathbf{Q}_t$ and $\overline{\mathbf{Q}}_t$; the storage requirement is $\mathcal{O}(n^2)$ in the shape/texture length, which may be quite large. The equivalent formulation $\mathbf{y}_t^* = [\alpha \mathbf{Y}_t \oplus (1-\alpha)[\mathbf{Y}_t \otimes [\mathbf{Y}_t^T \otimes \mathbf{Y}_s]]] \otimes \rho_s$ can be used, where ρ_s denotes the vector of source model parameters. The matrix applied to ρ_s does not depend on the expression and may be precomputed.

3.2 Similarity transform

The previous discussion addressed the reproduction of observed *non-rigid* variation. *Rigid* variation describes *scale*, *rotation* and *translation*, along with an appropriate rotation and shift of the colour palette to match skin tone. The alignment of the synthesised face to the head in the target scene must be sufficiently robust even in continuous video sequences, which is difficult because of the human sensitivity to the smallest inconsistencies in the positioning of key features such as eyes.

The procedure aims to align the landmark points of the synthesised face to those on the target head. For each landmark point i, let $\mathbf{x}_i$, $\mathbf{y}_i$ denote respectively the source and target locations of the point, and w_i a weighting summing to 1. If our objective is to orient the face to minimise the error measure $\sum_i w_i ||c\mathbf{R}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i||^2$, then this is the *absolute orientation problem*, whose solution may be given in terms of the singular value decomposition [9]:

$$\left\{ \begin{array}{rcl} \mathbf{R} & = & \mathbf{U}\mathbf{S}\mathbf{V}^T \\ \mathbf{t} & = & \boldsymbol{\mu}_y - c\mathbf{R}\boldsymbol{\mu}_x \\ c & = & \mathrm{trace}[\mathbf{D}\mathbf{S}]/\sigma_x^2 \end{array} \right.$$

where $\mathbf{U}\mathbf{D}\mathbf{V}^T$ is the SVD of $\mathbf{Z}$, and:

$$\left\{ \begin{array}{rcl} \boldsymbol{\mu}_x & = & \sum_i w_i \mathbf{x}_i \\ \boldsymbol{\mu}_y & = & \sum_i w_i \mathbf{y}_i \\ \sigma_x^2 & = & \sum_i w_i ||\mathbf{x}_i - \boldsymbol{\mu}_x||^2 \\ \mathbf{Z} & = & \sum_i (\mathbf{y}_i - \boldsymbol{\mu}_y)(\mathbf{x}_i - \boldsymbol{\mu}_x)^T \\ \mathbf{S} & = & \mathrm{diag}[1, 1, \ldots, 1, \mathrm{sign}(\mathbf{Z})] \end{array} \right.$$

We must ensure weights w_i are selected such that they describe the relative correlation of the motions of each of the landmark points with the *rigid motions of the skull*. For example, if the lower jaw receives significant weighting, the whole synthesised face would appear to shrink as the mouth opens, to compensate. Likewise, landmark points on the lips and eyelids should be given small or zero weighting - they may rise and fall quite independently to any overall head motion. Lending high weighting to various features in the upper face, including the corners of the eyes, the upper nose and temple, was found to produce visually acceptable results.

To determine the similarity transform to apply to the texture layer (corresponding to a rotation and shift of the colour palette so as to match the target image), the previous technique may be applied, with the set of pixel colour coordinates (R, G, B) in place of the set of vertex coordinates. Since it is less clear in this case what kind of weighting scheme should be applied, a simple uniform weighting may be substituted.

4. Experiments and Results

For the evaluation of the proposed methods, we ran each of the three methods on four seconds of speech from the AVOZES data corpus [5], swapping the speaker's face and/or head with two other subjects. Gaussian blurring of an alpha-mask was employed to smooth the boundaries of the synthesised faces with the target frames. The results give visual confirmation of the effect of varying the α parameter:

- $\alpha = 1$ - the direct transfer method - offers expression transfer very 'true' to the source sequence, but generates visually unacceptable artefacts.
- $\alpha = 0$ - the projection method - offers smoother and more convincing expressions, but with noticeably 'restrictive' motion caused by the hard constraint.
- $\alpha = 0.6$ - the weighted approach - balances the two effects and produces qualitatively better results.

Figure 1 shows the results of expression and face swapping on three sample images with two auxiliary models, using $\alpha = 0.6$.[3]

We also observe in the results some of the limitations of the linear approach to the problem. In cases where the constructed face is of a very different shape to that appearing in the target scene, some *feature duplication* can be observed. This occurs when the edge of the model face recedes back over the corresponding feature in the target head, leaving both sets of features visible. In the last row of Figure 1, two chins are visible. By analogy, we can also observe feature *subduction*, where the pasted face covers adjacent features not present in the tracked face area. These problems could potentially be overcome by applying a non-linear warp to some of the surrounding image so as to minimise the effect.

5. Conclusions

We have given a theoretical justification for a procedure for expression transfer between AAMs that generalises two well-known approaches. The procedure is simple enough to be applied in real-time and has been seen to produce video sequences that are smooth and seemingly acceptable. However, two remaining clues that the synthesised video sequences are 'fake' are particularly visible: the occasional phenomena of feature duplication and subduction. Such problems may in the

[3] Sample videos are available at http://users.rsise.anu.edu.au/∼roland/videos_icpr2010.zip.

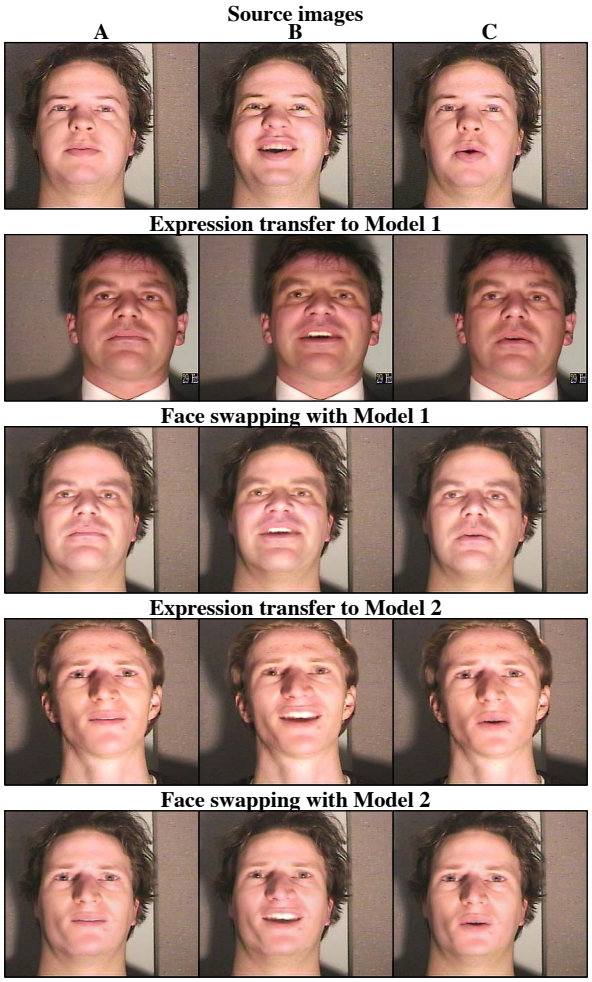

Figure 1. Expression/face swapping on three sample images from AVOZES

future be overcome with suitable non-linear extensions to the affine AAM transfer framework. Furthermore, we plan to perform perceptual experiments, in which human observers are asked to decide whether a given video is an original or synthesised one.

6. Appendix

Proof of MSE minimality of the projection method. $\mathbf{y}_t$ is constrained to lie within the range of $\mathbf{Y}_t$, thus $\mathbf{Q}_t\mathbf{y}_t = \mathbf{y}_t$ and thus by the first order condition:

$$\begin{aligned} \nabla_{\mathbf{y}_t}||\mathbf{y}_t - \mathbf{y}_s||^2\mathbf{Y}_t &= 2(\mathbf{y}_t^T + \mathbf{y}_s^T)\mathbf{Y}_t \\ \therefore \mathbf{Y}_t^T\mathbf{y}_t &= \mathbf{Y}_t^T\mathbf{y}_s \\ \mathbf{Q}_t\mathbf{y}_t &= \mathbf{Q}_t\mathbf{y}_s \\ \mathbf{y}_t &= \mathbf{Q}_t\mathbf{y}_s \end{aligned} \qquad \square$$

Proof of Theorem 1. From the first order condition, we obtain:

$$\begin{aligned} \nabla_{\mathbf{y}_t}C &= 2(\mathbf{y}_t - \mathbf{y}_s)^T + 2\varphi\mathbf{y}_t^T\overline{\mathbf{Q}}_t \\ \therefore \mathbf{y}_s &= [\mathbf{I} + \varphi\overline{\mathbf{Q}}_t]\mathbf{y}_t^* \end{aligned}$$

Projecting $\mathbf{y}_s$ into the complementary subspaces of $\mathbf{Q}_t$ and $\overline{\mathbf{Q}}_t$, we obtain:

$$\begin{aligned} \mathbf{Q}_t\mathbf{y}_s &= \mathbf{Q}_t\mathbf{y}_t^* \\ \overline{\mathbf{Q}}_t\mathbf{y}_s &= (1+\varphi)\overline{\mathbf{Q}}_t\mathbf{y}_t^* \end{aligned}$$

Merging the components, we have finally:

$$\begin{aligned} \mathbf{y}_t^* &= \mathbf{Q}_t\mathbf{y}_t^* + \overline{\mathbf{Q}}_t\mathbf{y}_t^* \\ &= \mathbf{Q}_t\mathbf{y}_s + [1/(1+\varphi)]\,\overline{\mathbf{Q}}\mathbf{y}_s \\ &= \left(\mathbf{Q}_t + \alpha\overline{\mathbf{Q}}_t\right)\mathbf{y}_s \end{aligned} \qquad \square$$

References

[1] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. Nayar. Face Swapping: Automatically Replacing Faces in Photographs. *ACM Trans. on Graphics (also Proc. of ACM SIGGRAPH)*, 27(3), Aug. 2008.

[2] V. Blanz, K. Scherbaum, T. Vetter, and H. Seidel. Exchanging Faces in Images. In *Proc. EUROGRAPHICS 2004*, pages 669–676, Grenoble, France, 2004.

[3] V. Blanz and T. Vetter. Face Recognition Based on Fitting a 3D Morphable Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, Sept. 2003.

[4] G. Edwards, C. Taylor, and T. Cootes. Interpreting Face Images Using Active Appearance Models. In *Proc. FG'98*, pages 300–305. IEEE, Apr. 1998.

[5] R. Goecke and B. Millar. The Audio-Video Australian English Speech Data Corpus AVOZES. In *Proc. ICSLP2004*, volume III, pages 2525–2528, Jeju, Korea, 2004.

[6] J. Saragih and R. Goecke. Learning AAM fitting through simulation. *Pattern Recognition*, 42(11):2628–2636, 2009.

[7] B. Theobald, I. Matthews, J. Cohn, and S. Boker. Real-time expression cloning using appearance models. In *Proc. ICMI'07*, pages 134–139, Nagoya, Japan, 2007.

[8] L. Trefethen and D. Bau. *Numerical linear algebra*. International Series in Natural Philosophy. SIAM: Society for Industrial and Applied Mathematics, Philadelphia, 1997.

[9] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 13(4):376–380, Apr. 1991.