Real-time Expression Cloning using Appearance Models

B. Theobald
School of Computing Sciences
University of East Anglia
Norwich, UK
bjt@cmp.uea.ac.uk

I. Matthews
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA
iainm@cs.cmu.edu

J.F. Cohn
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA
jeffcohn+@cs.cmu.edu

S. Boker
Department of Psychology
University of Virginia
Charlottesville, VA, USA
smb3u@virginia.edu

ABSTRACT

Active Appearance Models (AAMs) are generative parametric models commonly used to track, recognise and synthesise faces in images and video sequences. In this paper we describe a method for transferring dynamic facial gestures between subjects in real-time. The main advantages of our approach are that: 1) the mapping is computed automatically and does not require high-level semantic information describing facial expressions or visual speech gestures. 2) The mapping is simple and intuitive, allowing expressions to be transferred and rendered in real-time. 3) The mapped expression can be constrained to have the appearance of the target producing the expression, rather than the source expression imposed onto the target face. 4) Near-videorealistic talking faces for new subjects can be created without the cost of recording and processing a complete training corpus for each. Our system enables face-to-face interaction with an avatar driven by an AAM of an actual person in real-time and we show examples of arbitrary expressive speech frames cloned across different subjects.

Categories and Subject Descriptors

I.3.3 [**Computer Graphics**]: Picture/Image Generation;; I.3.7 [**Computer Graphics**]: Animation;; I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism;

General Terms

Human Factors, Measurement, Performance

Keywords

active appearance models, facial animation, expression cloning

1. INTRODUCTION

Realistic animation of human faces is challenging as the changes in the features of the face that we interpret as expressions are the product of a complex interaction between various anatomical layers, which include bone, muscle, subcutaneous fat, and skin. The problem is compounded by the fact that we are all expert at detecting and recognising facial expressions and so are sensitive to even the smallest discrepancies from normal behaviour. The central problem then is how best to approximate the intricacies of the face with sufficient detail such that expressions synthesised on a model look realistic?

Traditional facial animation approaches are graphics-based, where points on the surface of the face are represented as vertices in three-dimensions (3D) and the skin approximated by connecting the vertices to form a connected mesh. These mesh vertices are manipulated using time-varying parameters that influence the mesh geometry either directly, or using a physically-based approach [20]. Directly parameterised animation [17, 19] uses geometric interpolation between a collection of hand-crafted face models, known as morph-targets, where each is meticulously designed to be a faithful representation of a change in a particular aspect of the facial anatomy. The drawbacks of this approach are: 1) the morph-targets are usually designed by hand, which is time-consuming, 2) the morph-targets are designed for a particular mesh topology, so are not readily transferable across models. 3) The morph-targets generally are not independent, so care is required to ensure a valid facial expression results from any given combination of the morphs.

Indirectly parameterised models are designed to approximate the anatomical structure of the face, where animation parameters act on physical models, which in turn update the mesh geometry. A popular approach is Waters' pseudo-muscle model [26], where individual mesh vertices are displaced according to the relative vicinity of nearby muscle functions embedded within a mesh. Improved realism has been achieved by extending this approach to use physically-based methods [14, 22]. The limitations of physically-based animation are: 1) it is relatively computationally expensive compared to directly parameterised animation as the influence on each individual vertex must be computed as a function of each muscle. 2) While anatomical models are not tied to a particular mesh topology they must be manually

inserted in the mesh. Incorrectly embedding a muscle will produce unexpected results when the model is animated. 3) To prevent artifacts in the rendered mesh care is required to ensure that discontinuities at the boundaries of the regions of interest of the anatomical models are taken into account.

Image-based synthesis can produce animated sequences with a high degree of both static and dynamic realism. Typically animation is achieved either using a data-driven approach, where frames in an existing video sequence are reordered [4, 9], by morphing between static images representing key-frames in a video sequence [11], or by warping images using control parameters generated by trajectory synthesis [3, 10]. The main limitations of image-based animation are: 1) it is relatively computationally expensive compared with graphics-based systems, 2) animation is usually confined to only re-animating the face, and 3) transfer of speech and expression information between subjects is relatively difficult and somewhat constrained compared with graphics-based approaches.

Hybrid systems that animate both the geometry *and* the appearance of the face have been proposed. One approach is to construct an Active Appearance Model (AAM) [8] from a video sequence and use the model to re-animate a face speaking novel phrases [24]. A similar idea is to use a 3D morphable model (3DMM) to render existing faces displaying new expressions [2]. Alternatively, 3D point locations for a sparse set of points on a face can be recovered from a number of different views using photogrammetry. A dense geometric mesh can then be fitted to the recovered points and the images from each view blended to create view-dependent texture maps. Repeating the process for a number of expressions allows realistic sequences to be animated by interpolating the geometry and morphing the view-dependent images (across both view and time) [21]. Extending this idea to capture the reflectance field of the face allows for a change in illumination in the animated sequences [13]. These latter methods animate expressive sequences with a stunning degree of realism, but it is not immediately clear how such techniques can be extended to animate subtle facial gestures, such as those corresponding to arbitrary speech phrases.

Performance-driven facial animation transfers movements on the face of an actor to a model. An advantage of this over synthesis-based animation is the level of naturalness and realism that can be achieved. One method involves locating a few key feature points on the face of an actor and either interpolating the displacement of these feature points directly to mesh vertices [7, 12, 18, 27], or mapping the motion to an underlying physically-based model for synthesis [6]. In addition, expressions can be cloned from images of one subject to images of another by considering the change in both the geometry and appearance of the face [5, 15, 28]. However, such image-based methods are relatively computationally expensive and are usually limited in the range of expression that can be transferred. It is also difficult to manipulate the cloned expression, such as exaggerate or attenuate the degree of expressiveness, and the result is usually an imposition of the source expression on the target face. The 3DMM has been used to transfer expression information between faces, which offers a compromise between between graphics-based and image-based approaches [1]. While the results can look convincing, the main disadvantages are: 1) expressive information is not *mapped* to a target face. Rather changes in the facial features for one person are simply copied to another person. 2) The same inner-mouth is used for all subjects. 3) The algorithm is relatively computationally expensive and requires a collection of laser scans of faces. The lack of expression mapping in [1] was recently addressed in [25], where multi-linear models are constructed from a number of people speaking and displaying pre-specified facial expressions. The model is matched to new faces and expressions cloned on these new faces based on statistics learned from the training data. In particular the multi-linear model captures the variation due to identity, expression and speech independently. This approach was also approximated in 2D and shown to work with AAMs [16].

In this paper we describe a simple and intuitive mapping between AAMs for two or more people such that expressions on one face can be transferred to other faces. Our approach requires no high-level semantic information describing facial expressions, as is required in [16, 25], it allows the full face to be transferred, unlike [1] and operates at video frame-rate.

2. ACTIVE APPEARANCE MODELS

The *shape* of an AAM is defined by a 2D triangulated mesh and in particular the vertex locations of the mesh. Mathematically, the shape $\mathbf{s}$ of an AAM is defined as the concatenation of the x and y-coordinates of the n vertices that make up the mesh: $\mathbf{s} = (x_1, y_1, \ldots, x_n, y_n)^T$. A compact model that allows a linear variation in the shape is given by,

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i, \tag{1}$$

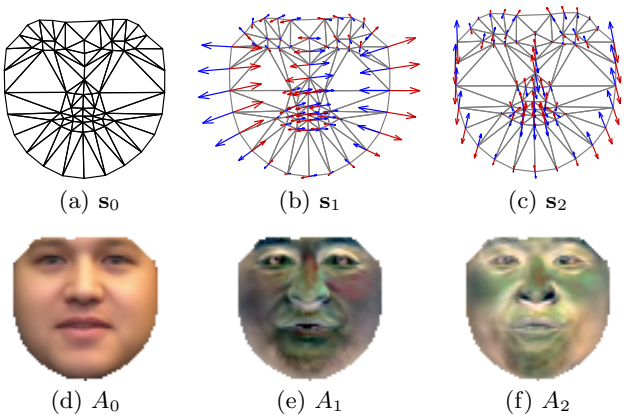
where the coefficients p_i are the shape parameters. Such a model is usually computed by applying principal component analysis (PCA) to a set of meshes hand-labelled in a corresponding set of images [8]. The base shape $\mathbf{s}_0$ is the mean shape and the vectors $\mathbf{s}_i$ are the (reshaped) eigenvectors corresponding to the m largest eigenvalues.

The *appearance* of the AAM is defined within the base mesh $\mathbf{s}_0$. Let $\mathbf{s}_0$ also denote the set of pixels $\mathbf{x} = (x, y)^{\mathrm{T}}$ that lie inside the base mesh $\mathbf{s}_0$. The appearance of the AAM is then an image $A(\mathbf{x})$ defined over the pixels $\mathbf{x} \in \mathbf{s}_0$. AAMs allow linear appearance variation. This means the appearance $A(\mathbf{x})$ can be expressed as a base appearance $A_0(\mathbf{x})$ plus a linear combination of l appearance images $A_i(\mathbf{x})$:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{l} \lambda_i A_i(\mathbf{x}) \qquad \forall\ \mathbf{x} \in \mathbf{s}_0, \tag{2}$$

where the coefficients λ_i are the appearance parameters. As with the shape, the base appearance A_0 and appearance images A_i are usually computed by applying PCA to the (shape normalised) training images [8]. The base appearance A_0 is the mean shape normalised image and the vectors A_i are the (reshaped) eigenvectors corresponding to the l largest eigenvalues. An example appearance model is shown in Figure 1.

A near-photorealistic image of a face is rendered using the AAM by first applying the shape parameters $\mathbf{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$, Equation (1) to generate the shape, $\mathbf{s}$, of the AAM, then applying the appearance parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l)^{\mathrm{T}}$ to generate the AAM image, $A(\mathbf{x})$. The final rendered image is created using a piece-wise affine warp to warp $A(\mathbf{x})$ from the base shape, $\mathbf{s}_0$, to the model-generated shape, $\mathbf{s}$.



(a) $\mathbf{s}_0$ (b) $\mathbf{s}_1$ (c) $\mathbf{s}_2$

(d) A_0 (e) A_1 (f) A_2

Figure 1: The shape (top row) and appearance (bottom row) of an AAM. Shown are the base shape (a) and base appearance (d), and the first two basis vectors of the respective models.

3. EXPRESSION CLONING USING AAMS

We now turn our attention to the central problem: automatically cloning gestures from a source face in a video sequence to any number of target faces. In particular, we are interested in mapping parameters between AAMs that describe person-specific speech and expression information. The models used here do not require control over the type or degree of expressiveness in the training data and we typically build our models from only 15–20 images per person.

3.1 Mapping Parameters Between Models

We propose a linear mapping that is intuitive given the nature of the vectors that span the shape and appearance space of the AAM. Each component of a shape vector is an offset from the mean shape (resp. appearance) and the vector itself represents the overall displacement that gives rise to a specific type of gesture — see Figure 1. For example, one vector might be responsible for opening and closing the mouth, while another might control eye blink, and so on. If the correspondence between models were one-to-one, we could simply apply the parameters for one model directly to the shape/appearance vectors of another (ignoring scale). However, it is extremely unlikely that the vectors will correspond between models in this way. Indeed it could be that a specific source of variation captured by a *single* basis vector for one model is represented as a *combination* of basis vectors for another model.

To map the meaning of the parameters from one model to another we compute the relationship between the basis vectors in the two model-spaces to determine the *combination* of vectors in the target space that produces the corresponding change in shape (or appearance) when moving along a *single* vector in the source space. As the basis vectors are unit length and can be constrained to lie in the same dimension Euclidean space when the models are built, the alignment of a source basis vector with the target vector-space is given simply by the inner products $< \mathbf{s}_i^s, \mathbf{s}_j^* >$ between a source vector and each of the target vectors. Thus, a vector (a displacement from the mean) in the source space is a weighted average of the vectors (displacements from the mean) in the target-space, and the weights are obtained from the inner-products. More formally, expressing Equation 1 in matrix form and including the mapping gives:

$$\mathbf{s}^* = \mathbf{s}_0^* + \mathbf{S}^* (\mathbf{R}\mathbf{p}_s), \quad (3)$$

where the columns of $\mathbf{S}^*$ are the basis vectors spanning the **target** space, $\mathbf{R}$ is a $q \times r$ matrix of inner products (the target space is of dimension q and the source of dimension r), and $\mathbf{p}_s$ are the parameters representing the expression in the **source** space. Each parameter *value* in the source space therefore maps to a parameter *vector* in the target space. Note: $\mathbf{R}$ does not depend on expression and can be pre-computed, so the cost of mapping an expression is only a matrix-vector product. Notice here we do not explicitly need to define anything about the facial expressions. This information is implicit from the basis vectors in the source and target spaces. Also note we are not concerned with the direction of the eigenvectors. For example, an increasingly positive value for a source parameter might, say, open the mouth, while the same action could be defined by an increasingly negative value in the target space. In this instance the inner product for that combination of vectors is negative (the displacements from the mean are largely in opposite directions), so the sign of the parameter value is flipped when the parameter is mapped. Another important consideration when mapping parameters is moving too far along the target vectors, which could generate implausible faces. An obvious example is the upper and lower lip boundaries intersecting. However, the parameters after mapping to the target space can be constrained with the limits of the original (target) training data (typically they must lie within $\pm 3\sigma$ from the mean), which will ensure only valid faces with the target appearance are generated by the mapping.

The underlying assumption of our approach is that the base shape and base appearance vectors for the source and target models represent similar expressions. Since these vectors are averages of a number of facial gestures this assumption usually holds as the average expression quickly converges, even for different faces — see Figure 2.

Figure 2: The mean appearance images for three individuals computed from approximately 15 images per person. In general the 'average expression' is a partially open mouth with upper teeth slightly visible.

3.2 Recovering Missing Components

The linear mapping described previously will undoubtedly lose information. If a source vector cannot be completely described by the target space the mapped expression/speech gesture will appear under-articulated compared with the original. However, we can determine beforehand what components of the source model cannot be described

by the target space and account for this in the target model. We proceed as follows:

1. Perturb the shape of the source by $+1\sigma$ along the i^{th} basis vector:

$$\delta_s^i = \mathbf{s}_i \sqrt{\gamma_i}$$

where γ_i is the variance captured by the i^{th} shape vector.

2. Map the source parameter to the target space and perturb the target vectors:

$$\delta_t = \mathbf{S}^* \left(\mathbf{R}\sqrt{\gamma_i}\right)$$

This is the displacement of the source space as best represented in the target space using the linear mapping.

3. Finally, the residual:

$$\delta_i = \delta_t - \delta_s^i$$

gives the component of the i^{th} source vector that cannot be reconstructed by the target model. This is repeated for each shape vector and multiples of these offsets (determined by the shape parameters) are added to the shape reconstructed on the target model.

$$\mathbf{s}^* \;=\; \mathbf{s}_0^* + \mathbf{S}^* \left(\mathbf{R}\mathbf{p}_s\right) + \delta_s \mathbf{p}_s. \qquad (4)$$

The appearance parameters are mapped using Equation 3, but we do not compensate for the appearance residuals [1].

4. RESULTS

To compare our mapping with related approaches we have implemented, in the framework of our AAM system, a similar method to those described in [1] and [15], where expression information for one individual is simply copied across to a target face(s). In the context of an AAM this is simply a substitution of the target mean shape and appearance into the source model, with appropriate constraints to ensure number of shape vertices and appearance pixels are the same for both subjects.

To obtain expressive data we filmed a source subject (a female) while speaking and displaying facial expressions. In real-time we track this source subject using an AAM and map the parameters to the target models (one each for a male and female subject), then render the resultant images. An (original) image of the source and the two targets is shown in Figure 3. Note: in the examples all of the subjects are smiling — this is so the reader can compare the reconstructed inner mouth in our cloned examples with the original. The inner mouth is an area of the face that is a particular problem for expression cloning systems as it undergoes significant variation (mouth can be open or closed, the teeth may or may not be visible, and the tongue also displays varying degrees of visibility). For this reason the inner mouth is often ignored when cloning expressions: [15] clone between expressions where the mouth is always closed, and [1] insert a generic set of teeth into all target faces. The advantage of an AAM is the variation of the inner mouth is also captured by the appearance vectors that span the appearance space.

Example facial gestures on the source identity cloned onto the two target models are shown in Figure 4. Simply swapping (i.e. no mapping) the expression information between faces creates artifacts on the cloned face (especially around the teeth), where some of the characteristics of the source identity are imposed onto the target face. Mapping the parameters from the source to the target model allows us to constrain the expression on the target face, so the characteristics of the target are retained — see Figure 4.

Figure 3: Example images of the source (center) and target (left and right) subjects used to test our expression cloning mapping.

5. LIMITATIONS AND EXTENSIONS

Our expression cloning system using AAMs has a similar requirement to image-based cloning [5, 28] and morphable model-based cloning [1] in that the source and target reference expressions must be similar. In our case this means the base shape, $\mathbf{s}_0$, and appearance, $A_0(\mathbf{x})$, must represent a similar expression. This is because the model parameters represent perturbations about the base shape/appearance, so applying the parameters to a different reference expression can produce artifacts in the rendered faces. However, in our experience the mean shape and appearance for different models tends to converge relatively quickly. Of course it is entirely possible to force the base shape and appearance for the target to be similar to the source model. There is no reason the base vectors need be their respective means. The images with the closest shape vertices and appearance pixels, either visually or in terms of the L_2-norm, could be selected as the reference expression and the basis vectors for the target model calculated with respect to these references.

A second limitation is the quality of the images reconstructed using the AAM. The images typically suffer spatial blur in the regions of the face that matter most (the eyes and the mouth) since these are the regions that least satisfy the underlying linearity assumption of PCA. For a real-time system the amount of processing that can be performed to improve the image quality is constrained. However, we are currently investigating multi-segment appearance models [23], extended to partition the appearance-space using the shape parameters. The appearance is then composed of a number of sub-models, which are easily combined to generate the appearance image.

6. SUMMARY

In this paper we have described techniques for transferring visual speech information and facial expressions between faces. An AAM is constructed for a source face and one or more target faces. The source face is tracked in a video sequence and the model parameters representing the face in each frame are mapped to the target face(s) and re-rendered, all in real-time.

The advantages of using AAMs are 1) the mapping is simple and intuitive. 2) The model can account for a high degree of variability in the images offering more flexibility than image-based approaches. 3) No semantic information regarding the expression is required as the expression is implicitly coded by the parameters of the model. 4) Near-videorealistic talking faces for new subjects can be created without the cost of recording and processing a complete training corpus for each. 5) The use of the model allows the mapped expression to be constrained, so it has, as far as possible, the appearance of the target producing the expression, rather than the source expression on the target face. We can also manipulate the parameters to exaggerate or attenuate the mapped expression, which is difficult with comparable image-based approaches. The system described in this paper has been used in subjective evaluation involving live face-to-face video conferencing and we have found people are unable to identify when they are speaking with re-render video or when they are speaking with a clone. These experiments will be the focus of a future publication.

Acknowledgements

The research described in this paper was supported in part by NSF Grant BCS–0527485, NSF Grant HSD–0(49)527444, and EPSRC Grant EP/D0490751.

7. REFERENCES

[1] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Eurographics*, pages 641–650, 2003.

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of SIGGRAPH*, pages 187–194, 1999.

[3] M. Brand. Voice puppetry. In *Proceedings of SIGGRAPH*, pages 21–28, 1999.

[4] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of SIGGRAPH*, pages 353–360, 1997.

[5] Y. Chang and T. Ezzat. Transferable videorealistic speech animation. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 29–31, 2005.

[6] B. Choe, H. Lee, and H. Ko. Performance-driven muscle-based facial animation. *Journal of Visualization and Computer Animation*, 11(2):67–79, May 2001.

[7] E. Chuang and C. Bregler. Performance driven facial animation using blendshape interpolation. Technical Report CS-TR-2002-02, Stanford University, April 2002.

[8] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.

[9] E. Cosatto and H. Graf. Sample-based synthesis of photorealistic talking heads. In *Proceedings of Computer Animation*, pages 103–110, Philadelphia, Pennsylvania, June 1998.

[10] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of SIGGRAPH*, pages 388–398, 2002.

[11] T. Ezzat and T. Poggio. Miketalk: A talking facial display based on morphing visemes. In *Proceedings of the Computer Animation Conference*, pages 96–103, Philadelphia, Pennsylvania, 1998.

[12] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin. Making faces. In *Proceedings of SIGGRAPH*, pages 55–66, 1998.

[13] T. Hawkins, A. Wenger, C. Tchou, A. Gardner, F. . Goransson, and P. Debevec. Animatable facial reflectance fields. In *Eurographics Symposium on Rendering*, June 2004.

[14] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Proceedings of SIGGRAPH*, pages 55–62, 1995.

[15] Z. Liu, Y. Shan, and Z. Zhang. Expressive expression mapping with ratio images. In *SIGGRAPH*, pages 271–276, 2001.

[16] I. Macedo, E. Vital Brazil, and L. Velho. Expression transfer between photographs through multilinear aam's. In *Brazilian Symposium on Computer Graphics and Image Processing*, pages 239–246, 2006.

[17] D. Massaro. *Perceiving Talking Faces*. The MIT Press, 1998.

[18] J. Noh and U. Neumann. Expression cloning. In *SIGGRAPH*, pages 277–288, 2001.

[19] F. Parke. Parametric models for facial animation. *Computer Graphics and Applications*, 2(9):61–68, 1982.

[20] F. Parke and K. Waters. *Computer Facial Animation*. A K Peters, 1996.

[21] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. Salesin. Synthesizing realistic facial expressions from photographs. In *Proceedings of SIGGRAPH*, pages 75–84, 1998.

[22] D. Terzopoulos and K. Waters. Physically-based facial modelling, analysis and animation. *Journal of Visualization and Computer Animation*, 1(2):73–80, 1990.

[23] B. Theobald, J. Bangham, I. Matthews, and G. Cawley. Visual speech synthesis using statistical models of shape and appearance. In *Proceedings of Auditory Visual Speech Processing*, pages 78–83, 2001.

[24] B. Theobald, J. Bangham, I. Matthews, and G. Cawley. Near-videorealistic synthetic talking faces: Implementation and evaluation. *Speech Communication*, 44:127–140, 2004.

[25] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. *ACM Transactions on Graphcs*, 24(3):426–433, 2005.

[26] K. Waters. A muscle model for animating three-dimensional facial expressions. In *Proceedings of SIGGRAPH*, pages 17–24, 1987.

[27] L. Williams. Performance driven facial animation. *Computer Graphics*, 24(2):235–242, 1990.

[28] Q. Zhang, Z. Liu, G. Quo, D. Terzopoulos, and H. Shum. Geometry-driven photorealistic facial expression synthesis. *IEEE Transactions on Visualization and Compuer Graphics*, 12(1):48–60, 2006.

Figure 4: Mapping facial gestures from a source model (left column) to two target faces: one is female (top row of each block) and the other male (bottom row of each block). Expressions are cloned by (A) substituting the mean shape and appearance of the target into the source space, (B) mapping the parameters using Equation 3, and (C) mapping the parameters and compensating for the residuals using Equation 4.