Re-mapping Animation Parameters between Multiple Types of Facial Model

D. Cosker, P. L. Rosin, and D. Marshall

School of Computer Science, Cardiff University, U.K.
D.P.Cosker,Paul.Rosin,Dave.Marshal@cs.cardiff.ac.uk

Abstract. In this paper we describe a method for re-mapping animation parameters between multiple types of facial model for performance driven animation. A facial performance can be analysed automatically in terms of a set of facial action trajectories using a modified appearance model with modes of variation encoding specific facial actions. These parameters can then be used to animate other appearance models, or 3D facial models. Thus, the animation parameters analysed from the video performance may be re-used to animate multiple types of facial model. We demonstrate the effectiveness of our approach by measuring its ability to successfully extract action-parameters from performances and by displaying frames from example animations.

1 Introduction and Overview

Facial animation is a popular area of research, and one with numerous challenges. Creating facial animations with a high-degree of static and dynamic realism is a difficult task due to the complexity of the face, and its capacity to subtly communicate different emotions. These are very difficult and time-consuming to reproduce by an animator. For this reason, research continues to progress in developing new facial animation methods and improving existing ones. One of the most popular facial animation methods today is expression mapping, also known as performance driven animation [7–9, 1]. In expression mapping, the face is used an an input device to animate a facial model. This is popular since it can potentially directly transfer subtle facial actions from the actors face onto the facial model. This method of animation can also greatly reduce animation production time.

A common theme in work on expression mapping is that facial parameters only map between specific types of facial model [6, 9, 8]. This paper addresses the issue of re-using facial animation parameters by re-mapping them between multiple types of facial model. Facial actions along with intensities may be identified from real video performances using computer vision, parameterised, and used to directly animate video-realistic appearance models with different identities. These same parameters may also be mapped directly to onto the morph-targets of a 3D facial model to produce 3D facial animation [5]. Thus the facial parameters may be re-used since they map onto more than one type of facial model.

Figure 1 gives an overview of our approach. This paper therefore makes the following contributions:

- An approach for expression mapping between different facial appearance models and also between facial appearance models and 3D facial models.
- An approach for extracting meaningful facial action parameters from video performances.
- An approach for creating appearance models with intuitive basis-vectors for use in animation.

Expression mapping between image based models has previously been considered by several authors, e.g. [6, 10]. However, in these studies expressions are only transferred between the same type of model, and not between e.g. an image based model and a 3D blend-shape model.

Another such system for image based expression transfer is presented by Zhang *et al* [9]. Our method differs from theirs in several ways. Firstly, our approach represents a persons facial performance as a set of meaningful action parameters, and facial expression transfer is based on applying these parameters to a different facial model. Zhang *et al* transfer expressions via a texture-from-shape algorithm, which calculates sub-facial texture regions on the target face based on transferred shape information from the input performance. Our approach also differs from that of Zhang *et al* in that whereas they incorporate multiple sub-facial models for each person in order to facilitate transfer – each offering a different set of basis vectors – our approach requires only a single set of basis-vectors per person, where these vectors represent information for the entire face.

Zalewski and Gong [8] describe a technique for extracting facial action parameters from real video and then using these to animate a 3D blend-shape facial model. However, their facial parameters concentrate on mapping only full facial expressions. In our work, more specific sub-facial actions may be mapped onto a 3D model as well as full expressions.

This paper is organised as follows. In Section 2 an overview of appearance model construction is given. In Section 3 we describe how to extract meaningful facial parameters from video performances using appearance models, and how to use these parameters to animate other appearance models, or 3D blend-shape facial models. In Section 4 we show animation results, and quantitatively evaluate our appearance model mapping technique. We give conclusions in Section 5.

2 Data Acquisition and Appearance Model Construction

We filmed a male participant using an interlaced digital video camera at 25 fps. Lighting was constant throughout each recording. The participant performed three different facial expressions: *happiness*, *sadness* and *disgust* (see Figure 2). We broke each expression down into a set of individual facial actions. We also added four more actions for individual eye-brow control (see Table 1).

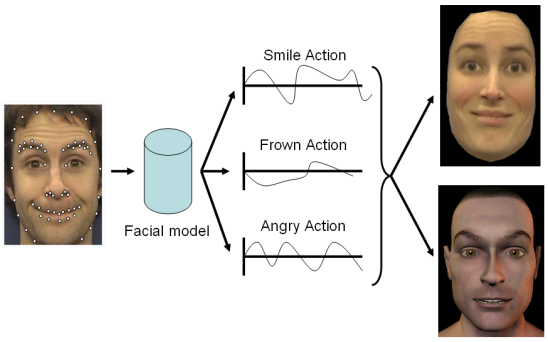

Fig. 1. Given a video performance we track facial features and project image and shape information into our facial model. This produces trajectories of facial action parameters which can be used to animate multiple types of facial model, namely appearance models and 3D facial models. Thus, animation parameters may be reused.

Expression	Actions
Happiness	(1) Surprised Forehead, (2) Smile
Sadness	(3) Sad Forehead, (4) Frown
Disgust	(5) Annoyed Forehead, (6) Nose Wrinkle
Miscellaneous	Left Eye-brow (7) Raise/ (8) Lower, Right Eye-brow (9) Raise/ (10) Lower

Table 1. Facial Actions.

We semi-automatically annotated the images in each video performance with 62 landmarks (see Figure 2) using the Downhill Simplex Minimisation (DSM) tracker described in [3]. We then constructed appearance model using the landmarked image data. A brief description of this procedure is now given. For further details, see [2].

We calculate the mean landmark shape vector $\bar{\mathbf{x}}$ and warp each image in the training set to this vector from its original landmark shape $\mathbf{x}$. This provides a shape-free image training set. Performing PCA on this set of image vectors gives $\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g\mathbf{b}_g$ where $\mathbf{g}$ is a texture vector, $\mathbf{P}_g$ are the eigenvectors of the distribution of $\mathbf{g}$, and $\mathbf{b}_g$ is a vector of weights on $\mathbf{P}_g$. Performing PCA on the training set of shape vectors gives the model $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_x\mathbf{b}_x$ where $\mathbf{P}_x$ are the eigenvectors of the distribution of $\mathbf{x}$, and $\mathbf{b}_x$ is a vector of weights on $\mathbf{P}_x$.

We now represent the training set as a distribution of joint shape ($\mathbf{b}_x$) and texture ($\mathbf{b}_g$) weight vectors. Performing PCA on this distribution produces a model where $\mathbf{x}$ and $\mathbf{g}$ may be represented as functions of an appearance parameter $\mathbf{c}$. We write $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_x\mathbf{W}^{-1}\mathbf{Q}_x\mathbf{c}$, and $\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g\mathbf{Q}_g\mathbf{c}$. Here, $\mathbf{Q}_x$ and $\mathbf{Q}_g$ are respective shape and texture parts of the eigenvectors $\mathbf{Q}$ - these eigenvectors

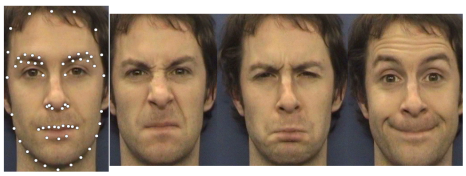

Fig. 2. Example landmark placement and participant facial expressions for *Disgust*, *Sadness* and *Happiness*.

Fig. 3. First four modes of variation for our male participant. Note that the modes encode combinations of facial actions.

belonging to the joint distribution of shape and texture weights. The elements of $\mathbf{c}$ are weights on the basis vectors of $\mathbf{Q}$. Each vector in $\mathbf{Q}$ describes type of facial variation. Figure 3 shows the first four modes of variation for the male participant.

3 Expression Mapping

We first describe how we create a new appearance model with parameters for specific facial actions. We then describe how these parameters can be used to animate other other appearance models, or 3D facial models with pre-defined morph-targets.

3.1 Creating Appearance Models with Action Specific Modes

We aim to build a new model with modes of variation controlling the actions in Table 1. One way to achieve this is to create individual appearance models for different sub-facial regions [3]. Since the highest mode of variation should capture the largest proportion of major texture and shape change, then in the case of e.g. modelling the lower part of the face given only images of a smile, then the highest mode of variation should provide a good approximation of that smile. By applying this rule to multiple facial regions we can obtain a set of modes over several sub-facial appearance models which provide our desired modes. We have previously shown this to be the case on several occasions [3]. However, managing multiple sub-facial appearance models becomes cumbersome, and blending these together can produce visual artefacts.

Here we provide a solution which provides all the benefits of using multiple sub-facial appearance models in a single facial appearance model. We break the face into four regions where actions 1, 3 and 5 belong to a forehead region (R_1), actions 2, 3 and 6 belong to a lower face region (R_2), actions 7 and 8 belong to a left eyebrow region (R_3) and actions 9 and 10 belong to a right eyebrow region (R_4). Let $G = (\mathbf{g}_1, \ldots, \mathbf{g}_N)$ be the training set of N shape free facial images. For each region we create a new set of images $R_j^G = (\mathbf{r}_1^G, \ldots, \mathbf{r}_N^G)$. In this set, $\mathbf{r}_i^G$ is constructed by piece-wise affine warping a region from image $\mathbf{g}_i$ over the mean image $\bar{\mathbf{g}}$. The boundaries between the superimposed image region and the mean image are linearly blended using an averaging filter. This removes any obvious joins. Figure 4 defines our four different facial regions, gives example images from each region, and illustrates construction of an artificial image. Images shown in this Figure are shape-free.

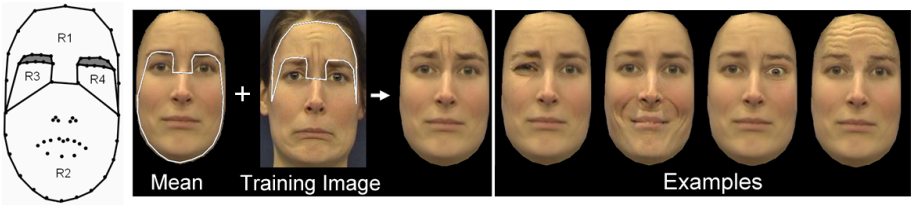

Fig. 4. (Left) There are four facial regions. The grey areas are shared by regions R_1, R_3 and R_4. (Middle) An artificial image is constructed by warping a region from a training image over the mean image. (Right) Example artificial images for actions (Left to Right) 8, 2, 9 and 10.

We now have a new training set of shape-free images $G' = (R_1^G, R_2^G, R_3^G, R_4^G)$ consisting of $4N$ artificial training images. The next task is to create a corresponding training set of artificial shape vectors. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ be the training set of N shape vectors. Again, we define a new set of vectors for each region $R_j^X = (\mathbf{r}_1^X, \ldots, \mathbf{r}_N^X)$. A vector $\mathbf{r}_i^X$ is constructed by calculating offsets between $\mathbf{x}_i$ and $\bar{\mathbf{x}}$ in a specific region, and then adding these to $\bar{\mathbf{x}}$. Figure 5 shows example training vectors superimposed for each region.

Fig. 5. (Left to Right) Superimposed shape vectors for regions R_1, R_2, R_3 and R_4.

We now have a new training set of shape vectors $X' = (R_1^X, R_2^X, R_3^X, R_4^X)$ consisting of $4N$ vectors. Performing PCA on X' and G' we have the new models $\mathbf{x}' = \bar{\mathbf{x}}' + \mathbf{P}'_x \mathbf{W}_x^{-1'} \mathbf{b}'_x$ and $\mathbf{g}' = \bar{\mathbf{g}}' + \mathbf{P}'_g \mathbf{b}'_g$.

A new AAM can be constructed from X' and G', where joint shape and texture information may be represented as a functions of an appearance parameter $\mathbf{c}'$. We define $\mathbf{v}$ as the concatenation of $\mathbf{x}'$ and $\mathbf{g}'$, and write our model as $\mathbf{v} = \bar{\mathbf{v}} + \mathbf{Q}'\mathbf{c}'$, where $\mathbf{Q}'$ are the eigenvectors of the joint artificial shape and texture distribution, and $\bar{\mathbf{v}}$ is the mean concatenated artificial shape and texture vector. Figure 6 shows the first four modes of appearance model variation for each participant, along with selected vector distributions for elements of $\mathbf{c}'$. Note that the modes of this model represent more localised facial variations than in the previous appearance model, where modes represent combinations of several facial actions at once (Figure 3). Thus this new representation allows us to parametrically control individual facial regions. Also note in this Figure how distributions of appearance parameters (red dots) representing specific facial regions are orthogonal to each other when viewed in this lower-dimensional form.

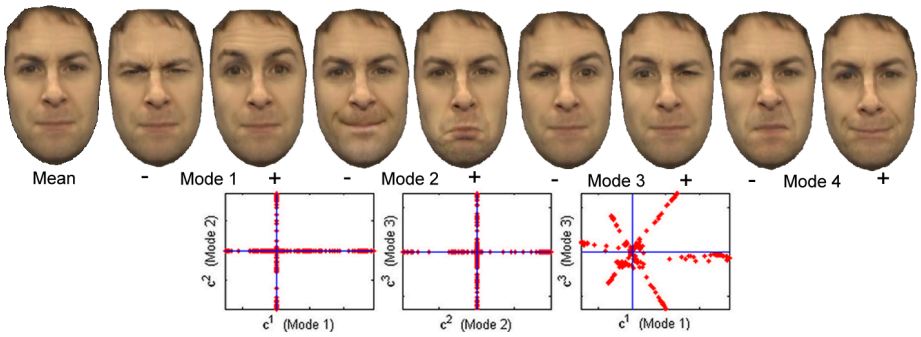

Fig. 6. Modes of appearance variation for the new appearance model, along with appearance parameter distributions visualised in a low dimensional space. Note how these modes capture localised facial variations, as opposed to the standard appearance model shown in Figure 3.

3.2 Mapping Performances Between Different Appearance Models

When appearance models are constructed for different people, the modes of variation of both models will in nearly all cases encode different types of facial variation, i.e. there will be no one-to-one mappings between the modes of variation in both models with respect to the specific facial actions they contain.

The technique described in the previous section to a great extent allows us to control what variations will be encoded in a particular mode. Therefore, appearance models can be constructed for two different people, and the modes

can be biased to encode our desired variations. Given a new facial performance from a person, it can be analysed in terms of the weights it produces on a specific set of modes. This forms a set of continuous parameter trajectories which can be mapped on the modes of variation of a second appearance model. This is how we achieve expression mapping between different appearance models in this work.

We identified modes in the male participants new appearance model relating to the 10 actions specified in Table 1. We then fit this model to a new video performance of the participant. Figure 7 demonstrates the 10 selected action modes along with example action-mode trajectories resulting from fitting the appearance model to the video. Note that some trajectories are given negative values. This relates to the fact that in order to produce the desired action on that particular mode, a negative value is required (this is also shown this way for clarity – if all trajectories were given positive values the figure would be far less clear).

The trajectories are normalised between -1 and 1 by dividing through by their maximum or minimum value. Just by examining the trajectories in Figure 7 it is easy to imagine what the participants video performance would have looked like. This is an advantage of having such an intuitive parameter representation.

We next recorded a new participant using the same set-up described in Section 2. We constructed a new appearance model for this person (using the approach described in Section 3.1), and identified modes for the same 10 actions (see Figure 7). Limits on the values of these modes relate to the maximum and minimum mode-weight values recorded from the training set. Now, given a set of action-mode trajectories from the male participant, an animation for the female participant can be produced by applying these to the corresponding action modes.

3.3 Mapping Between Appearance Models and 3D Facial Models

The model representation described gives us a clear set of facial action trajectories for a persons performance. The next aim is to use these to animate a morph-target based 3D facial model. In such a model, we define a set of facial actions by their peak expression. These act as the morph-targets, and we represent the intensity of these using a set of weights with values between 0 and 1. In practical terms, our morph-targets are just a set of 3D vertex positions making up the desired facial expression. The magnitude of a morph-target is its linear displacement from a neutral expression. Any facial expression in our 3D facial model can therefore be represented as

$$E = N + \sum_{i=1}^{n}((w(i)m(i)) - N) \tag{1}$$

where N is the neutral expression, w is a morph-target weight with a value between 0 and 1, m is a morph-target, and n is the number of morph-targets.

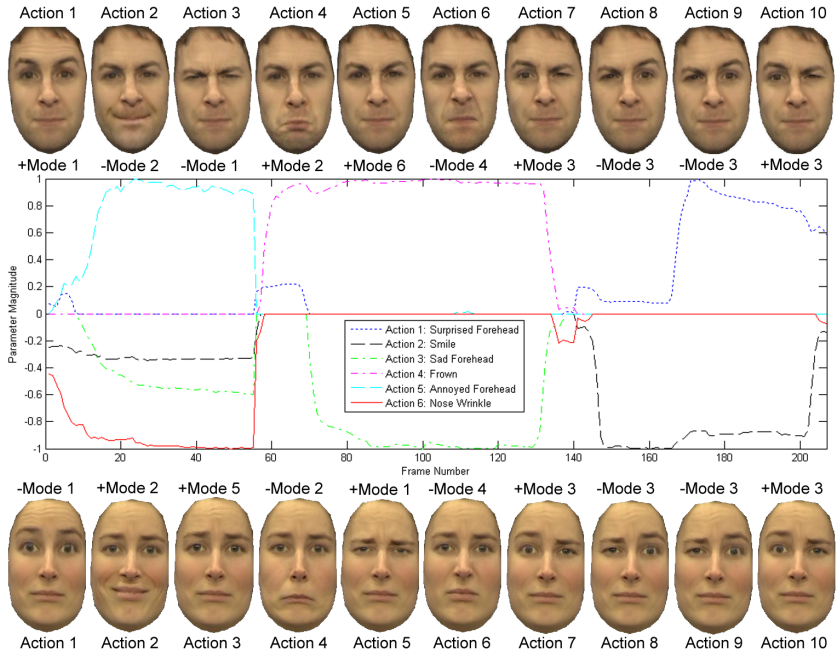

Fig. 7. (Top) Modes for the male participant relating to the actions specified in Table 1. (Middle) A video performance of the male participant represented as a set of continuous action-mode trajectories. (Bottom) Modes for the female participant relating to the actions specified in Table 1. Animations of this persons face can be created by applying the male action-mode trajectories.

Figure 8 shows the neutral expression and 6 morph-targets for our 3D male facial model.

New animations using this model can be created by fitting a persons appearance model to a facial performance, representing this performance as a set of action-mode trajectories, setting all these values to positive, and using these as values for w.

4 Results

In this Section we first demonstrate animation results before examining our approach from a quantitative perspective.

Figure 9 demonstrates expression mapping between our male participant, the new appearance model of our female participant, and our 3D morph-target based model. In our animations, expression transitions are smooth and in perfect synchronisation. We did not smooth the action-mode trajectories captured from our male participant before transfer onto the other facial models, and found that this did not degrade resulting animations. In fact, small changes in the action

Fig. 8. 3D facial model expressions. (Top-row left to right) Neutral, Surprised-forehead (Action 1), Smile (Action 2), Sad-forehead (Action 3), (bottom-row left to right) Frown (Action 4), Annoyed-forehead (Action 5) and Nose-wrinkle (Action 6).

trajectories often add to the realism of the animations, as these can appear as subtle facial nuances.

For the next part of our evaluation we investigated how well our new method for constructing appearance models encodes isolated facial variations. This is important since it is related to how well our model can recognise facial actions given a facial performance. If the model is poor at representing actions in individual modes, then the action-mode trajectories will be poor representations of the persons performance and the resulting performance driven animation will be less accurate.

Note that in a model with separate appearance models for different sub-facial regions, variations would be entirely confined to a certain region of the face. The following test may therefore also be considered an indirect comparison with a facial model of this kind.

We varied the weight on each action mode of our male appearance model up to its positive or negative limit, produced corresponding facial shape and texture, and then subtracted the mean shape and texture. The result is shown in Figure 10. The figure shows how major variations do occur in isolated regions. It also shows that some minor variations also simultaneously occur in other regions. The visual result of this in animations is negligible, and for practical animation purposes it may therefore be said that this new model has comparable performance to a model with separate local appearance models. In fact, the appearance of small residual variations in other parts of the face may in some ways be seen as an advantage over a model with local sub-facial appearance models, since it may be considered unrealistic in the first place to assume the affect of facial actions is localised to a specific facial region.

We further investigated how well our single action-modes encode our separate facial actions using a numerical measure. It is unlikely, even in a model with local facial appearance models, that an entire facial action would be perfectly encoded in a single mode of variation. This would assume that facial actions are linear in motion, when they are not. It is more likely that a single mode will encode

Fig. 9. Mapping expressions from our male participant onto our female participant and a 3D morph-target based facial model.

a very large proportion of the actions variation, while the rest of the variation will be spread over a small set of other modes. In this next test, we only aim to measure how distinct each of our 10 action-modes are from one another.

We took 10 images from the training set formed in Section 3.1 corresponding to 10 action specific images. Each image consisted of the mean face image, overlaid with a region specific image representing a specific facial action. For example, the image representing the smile action consisted of the mean face overlaid just on region R_2 with a smile image. We formed 10 corresponding action-specific shape vectors in a similar manner.

Projecting this information into our appearance model results in 10 appearance parameter vectors. We can measure the exclusivity of an action with respect to a mode by measuring the orthogonality of these vectors. We take the following measure adapted from [4] where $\mathbf{c}$ is an appearance parameter

$$M(\mathbf{c}_i, \mathbf{c}_j) = \frac{(\mathbf{c}_i \cdot \mathbf{c}_j)^2}{(\mathbf{c}_j \cdot \mathbf{c}_j)(\mathbf{c}_i \cdot \mathbf{c}_i)} \tag{2}$$

This returns a value between 0 and 1, where 0 indicates that the vectors are orthogonal. Table 2 compares the orthogonality of our actions. It can be seen from this result that a great many of the actions are orthogonal. Some actions have a low orthogonality. However, this is because these actions produce

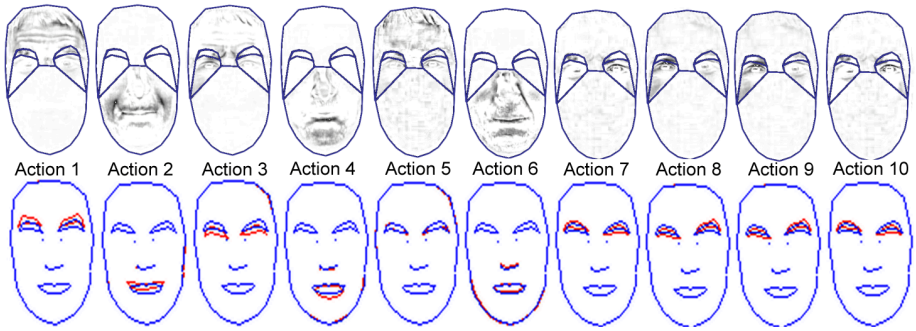

Fig. 10. Differences (for the male participant) between texture and shape vectors produced by our actions and their respective means.

	1	2	3	4	5	6	7	8	9	10
1	1	0.153	0.009	0.466	0.205	0	0	0	0.866	0.249
2	0.153	1	0.769	0.006	0.011	0	0	0	0.207	0.004
3	0.009	0.769	1	0.187	0.114	0	0	0	0	0.041
4	0.466	0.006	0.187	1	0.899	0	0	0	0.143	0.019
5	0.205	0.011	0.114	0.899	1	0	0	0	0.009	0.197
6	0	0	0	0	0	1	0.363	0.157	0	0
7	0	0	0	0	0	0.363	1	0.243	0	0
8	0	0	0	0	0	0.157	0.243	1	0	0
9	0.866	0.207	0	0.143	0.009	0	0	0	1	0.565
10	0.249	0.004	0.0410	0.019	0.197	0	0	0	0.565	1

Table 2. Orthogonality between Facial Actions 1 to 10 for our Male Participant.

variations in the same facial region which can produce changes in the same appearance mode.

In summary, these results show that our method for creating appearance models with action specific modes is successful, and is therefore well suited to producing performance driven animation in the way we have described. The results show that the modes of variation in our models successfully capture facial variations associated with actions when applied to video performances. This means that it is accurate enough to reliably transfer performances onto other facial models.

5 Conclusions and Future Work

We have presented a method for measuring facial actions from a persons performance and mapping this performance onto different types of facial model. Specifically, this mapping is between different appearance models, and between

appearance models and 3D morph-target based facial models. We have described how to construct appearance models with action-specific modes in order to record facial actions and represent them as continuous trajectories. We have also shown how these parameters are used for animating the models with respect to different facial actions.

By numerically measuring the orthogonality of our action modes, we have successfully demonstrated that they are well suited to accurately capturing facial actions. This therefore demonstrates the models suitability for performance driven facial animation applications.

One issue that we currently do not address is the transfer of inner mouth detail for e.g. smiles. This would be an interesting experiment, and it may be the case that further basis vector would be required to include this variation. We also do not consider the re-inclusion of head pose variation, and this would be the topic of future work. However, one solution which would part-way address this would be to reinsert the facial animations back into the training footage in a similar way to that described in [3].

References

1. V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Proc. of EUROGRAPHICS*, 2003.
2. T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. PAMI*, 23(6):681–684, 2001.
3. D. Cosker. Animation of a hierarchical image based facial model and perceptual analysis of visual speech. *PhD Thesis, School of Computer Science, Cardiff University*, 2006.
4. P. Gader and M. Khabou. Automatic feature generation for handwritten digit recognition. *IEEE Trans. PAMI*, 18(12):1256–1261, 1996.
5. P. Joshu, W. Tien, M. Desbrun, and F. Pighin. Learning controls for blend shape based realistic facial animation. In *Proc. of Eurographics/SIGGRAPH Symposium on Computer Animation*, 2003.
6. D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3):426–433, 2005.
7. L. Williams. Performance driven facial animation. *Computer Graphics*, 24(4):235 – 242, 1990.
8. L. Zalewski and S. Gong. 2d statistical models of facial expressions for realistic 3d avatar animation. In *Proc of IEEE Computer Vision and Pattern Recognition*, volume 2, pages 217 – 222, 2005.
9. Q. Zhang, Z. Liu, B. Guo, D. Terzopoulos, and H. Shum. Geometry-driven photorealistic facial expression synthesis. *IEEE Trans. Visualisation and Computer Graphics*, 12(1):48 – 60, 2006.
10. Z.Liu, Y. Shan, and Z. Zhang. Expressive expression mapping with ratio images. In *Proc. of SIGGRAPH*, pages 271–276, 2001.