Towards Automatic Performance Driven Animation Between Multiple Types of Facial Model

D. Cosker

School of Computer Science

University of Bath

D.P.Cosker@cs.bath.ac.uk

R. Borkett, D. Marshall and P. L. Rosin

School of Computer Science

Cardiff University

kbb@multiplay.co.uk, {*Dave.Marshall, Paul.Rosin*}*@cs.cf.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 in terms of a set of facial action parameter trajectories using a modified appearance model with modes of variation encoding specific facial actions which we can pre-define.

These parameters can then be used to animate other modified appearance models or 3D morph-target based 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. We also demonstrate its potential use in fully automatic performance driven animation applications.

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 [21, 22, 23, 3, 17]. In this approach, 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 actor's 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 [20, 23, 22, 17], and are not

general enough so that measured facial parameters may be used to animate several types of facial model (e.g. image based or 3D morph-target based). In some respects this is due to the fact that the mapping technique employed is tailored to work primarily between two specific classes of model (the input model used for analysis being regarded as one of these models). This paper addresses the issue of re-using facial animation parameters by incorporating an approach that allows them to be re-mapped between multiple types of facial model. This is made possible through the analysis of facial performances using specially created appearance models with modes of variation which are specifically tuned to be highly orthogonal with respect to pre-defined facial actions. These in turn provide an intuitive and meaningful representation of the performances facial actions over the performance.

We demonstrate how different facial actions – along with intensities – may be identified from real video performances, parameterised, and used to directly animate appearance models with different identities. These same parameters may also be mapped directly onto the morph-targets of a 3D facial model to produce 3D facial animation [12]. Thus the facial parameters analysed demonstrate a degree of model independence.

As well as describing our performance driven animation approach, we also consider how our methods may be employed in a fully automatic framework whereby output facial animations may be automatically created by recording a subject using a standard 2D video camera.

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 using appearance models.

- An approach for creating appearance models with intuitive basis-vectors for use in animation.

Active Appearance Models (AAMs) [6] have been found useful in many applications from e.g. tracking [15, 11] to animation [8, 18]. However, it is generally the case that the modes of variation in an appearance model (and indeed any PCA based facial model) do not correspond to independent facial actions – they typically encode mixtures of different facial actions depending on the variation contained in the training set (see Section 2). This means that when *fitted* to a new performance via tracking, the weights generated on different modes of variation over time do not encode information related to a unique facial action – instead the trajectories will describe the behaviors of mixtures of facial actions, and/or facial actions not related to a unique mode of variation. This makes it difficult to use standard PCA based facial models for facial action analysis through examination of the weights produced on individual modes of variation.

Expression mapping between image based models has previously been considered by several authors, e.g. [20, 14]. 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.

Using appearance models to transfer expressions has advantages over techniques such as *ratio images* [14] and 3D morphable models [3] since in these cases inner mouth detail is not present in the model and has to be artificially

created. This advantage is demonstrated by De la Torre and Black [13]. The difference between our approach and theirs is that we construct separate modified appearance models for each person and then produce our mappings, whereas De la Torre and Black define mappings beforehand by manually selecting corresponding frames from two facial data sets, and then perform Dynamic Coupled Component Analysis on this joint data set.

Another such system for image based expression transfer is presented by Zhang *et al.* [23]. Our method differs from theirs in several ways. Firstly, our approach represents a person's 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.

Chuang and Bregler [5] describe a method to map between a bi-linear image based model and a 3D morph-target model. Image based performances are represented as a set of weights on a set of image based key-frames, and then transfered to 3D morph-targets with similar facial configurations. The main difference between this work and our work is that our model can also be used to extract meaningful and *distinct* action trajectories over a performance during video analysis. The same may also be carried out using Chuang and Breglers

model. However, since their key-frames are not designed to be orthogonal with respect to certain actions, then the action trajectories analysed from video will be less distinct and meaningful.

Chang and Ezzat describe a method to transfer visual speech between realistic 2D morphable models. However, it is unclear how their approach could be applied to transferring facial expressions between models [4].

Zalewski and Gong [22] 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 consider how these mapping techniques may be incorporated into a fully automatic animation system. In Section 5 we show animation results, and quantitatively evaluate our appearance model mapping technique. We give conclusions in Section 6.

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

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.

three different facial expressions containing a combination of different facial movements (see Figure 2). The first expression contained a lowered brow, a nose-wrinkle, and a tightened mouth. The second expression contained a frown and a *worried* or *sad* forehead. The third expression contained a brow raise, a widening of the eyes and a smile. Note that these individual components can be reconstituted to form a range of recognisable stereotypical facial expressions by following the guidelines described in the Facial Action Coding System (FACS) [9]. FACS describes the full range of possible facial actions – or Action Units (AUs) – that can be produced by a person. It therefore defines a standard with which to measure and record facial behaviour. Using the guidelines described in FACS, a stereotypical *happiness* expression can be approximated from the smile component, a *sadness* expression using the frown and sad forehead movement, a *disgust* expression from the nose-wrinkle, a *fear* expression from the raised eyebrow, and an *anger* expression from the lowered-brow.

We broke each expression down into a set of individual facial actions. We also identified and labeled four more actions responsible for individual eye-brow control (see Table 1).

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 [7]. We then constructed an appearance model using the land-marked image data. A brief description of this procedure is now given. For further details, see [6].

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. Separately performing PCA on this set of image vectors and the set of shape vectors gives

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{1}$$

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_x \mathbf{b}_x \tag{2}$$

where $\mathbf{g}$ is a texture vector, $\mathbf{P}_g$ are the eigenvectors of the distribution of $\mathbf{g}$, $\mathbf{b}_g$ is a vector of weights on $\mathbf{P}_g$, $\mathbf{P}_x$ 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 function of an appearance parameter $\mathbf{c}$. We write this as

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_x \mathbf{W}^{-1} \mathbf{Q}_x \mathbf{c} \tag{3}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{Q}_g \mathbf{c} \tag{4}$$

Here, $\mathbf{Q}_x$ and $\mathbf{Q}_g$ are respective shape and texture parts of the eigenvectors $\mathbf{Q}$ - these eigenvectors 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 **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 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 [7]. Since the highest mode of variation should capture the largest proportion of texture and shape change, then in the case of, for example, 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 [7]. However, managing multiple sub-facial appearance models becomes cumbersome, and blending these together can produce visual artefacts.

In this paper we describe an alternative approach, and provide a solution which offers all the benefits of using multiple sub-facial appearance models in a single facial appearance model. In this model we attempt to construct modes of variation which are as orthogonal as possible with respect to our desired indi-

vidual facial actions. This allows us to then analyse a performance by tracking a video with the model and also animate the model intuitively by varying weights on the modes.

Previous work by Costen *et al.* [16] considers the extraction of *functional subspaces* from appearance models, such as identity, expression, pose and lighting. The question of orthogonality is also an issue in this work – one aim being to ensure identity information is encoded on a separate set of modes than the other subspaces. The major difference in our work is that we attempt to orthogonalise local variations within the face as opposed to global variations across the entire face.

We break the face into four regions where actions 1, 3 and 5 belong to a forehead region (R_1), actions 2, 4 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 $R_j^G = (\mathbf{r}_1^G, \ldots, \mathbf{r}_N^G)$, where $j = \{1, 2, 3, 4\}$ depending on which facial region the image set represents. In this set, $\mathbf{r}_i^G$ is constructed by piece-wise affine warping that part of the image $\mathbf{g}_i$ corresponding to region j on top of 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.

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 corre-

sponding 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.

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

$$\mathbf{x}' = \bar{\mathbf{x}}' + \mathbf{P}'_x \mathbf{W}_x^{-1'} \mathbf{b}'_x \tag{5}$$

$$\mathbf{g}' = \bar{\mathbf{g}}' + \mathbf{P}'_g \mathbf{b}'_g. \tag{6}$$

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 $\mathbf{b}'_x$ and $\mathbf{b}'_g$, and write our model as

$$\mathbf{v} = \bar{\mathbf{v}} + \mathbf{Q}'\mathbf{c}' \tag{7}$$

where $\mathbf{Q}'$ are the eigenvectors of the joint artificial shape and texture parameter distribution, and $\bar{\mathbf{v}}$ is the mean concatenated artificial shape and texture parameter. Figure 6 shows the first four modes of appearance model variation for our male 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 (see 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 training vectors

are highly orthogonal to each other when projected onto modes 1 and 2, or 2 and 3. This illustrates how the variations along these modes are capturing distinct facial actions. Projections onto modes 1 and 3 co-vary along the modes of the appearance model (horizontal and vertical axis), indicating that facial variations produced by these parameters may have similarities This is confirmed by noting the eyebrow variation produced both by mode 1 and mode 3.

3.2 Mapping Performances Between Different Appearance Models

When appearance models are constructed for different people in the standard way, 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 mapping between the modes of variation in both models with respect to the specific facial actions they contain. Also the modes of variation in a standard appearance model may contain combined facial variations that we would rather were contained in separate modes.

The technique described in this paper to a great extent allows us to control what variations will be encoded in a particular mode. This allows us to potentially avoid the issue of having multiple desired variations within one mode. Similarly, appearance models can then be constructed for two different people, and the modes can be biased to encode our desired variations. However, it should be noted that although we can help control what variations the modes will contain, we cannot guarantee beforehand *which* mode will contain this variation. For example, we cannot be sure that mode 1 in two different models will control the same facial action. At the present time we perform this model mode

alignment step manually.

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 parameter trajectories that can be mapped onto the modes of variation of a second appearance model with modes for the same actions. This is how we achieve expression mapping between different appearance models in this work.

We identified modes in the male participant's 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 direction along the mode that produces the measured action, e.g. action 6 is created by applying negative values to mode 4.

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 participant's video performance would have looked like. This is an advantage of having such an intuitive parameter representation.

We next recorded a new, female 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 as shown in Figure 8. Limits on the values of these modes relate to the maximum and minimum 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. Since the weights transferred between the corresponding action-modes of two modes are scaled by their maximum or minimum values, then the observed extrema of an expression in one person will relate to the same observed extrema in the other person. For example, if a full-smile is at the limit of mode 2 of action 2 in the male participant model, and a full smile is also at the limit of mode 2 in the female participant, then a full smile in one model will map to a full smile in the other model, and vice-versa.

Note in Figure 7 that mode trajectories for certain actions in one model may need to be inverted before being applied to a mode with a similar action in another model. For example, in the male participant appearance model a positive value in mode 1 relates to action one, while a negative value on mode 1 relates to the same action in the female model. The need to invert values in this way can easily be identified during the model mode alignment step.

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 person's 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

$$E = N + \sum_{i=1}^{n}((m(i) - N)w(i)) \tag{8}$$

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. Figure 9 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 person's 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 Automatic Video Analysis

As mentioned in Section 3.2, analysis of a video performance in terms of a set of action mode trajectories requires that the appearance model for a person is *fitted* to a new video performance, i.e. the facial performance is tracked using the appearance model. How to achieve this tracking is a subject of much research interest. Typical solutions are either to manually annotate a new video performance with landmarks at key facial positions, or to incorporate semi-automatic landmark placement algorithms. The DSM technique incorporated in Section 2 of this paper is one such example of semi-automatic landmark placement. Other examples include the Active Appearance Model tracking approaches described in e.g. [6] or [15].

In this Section we consider an alternative approach to estimating the facial

action parameters. Instead of tracking the facial performance with an appearance model, we extract some low-level features directly from the video sequence and use these in a facial action classifier. The output of the classifier is a level of confidence that a certain facial action is present in the current frame of the video sequence. This can be translated as a value between zero and 1 for a facial action, and then mapped either onto a specific mode of variation (e.g. an appearance model with action-specific models (see Section 3.1)) or a morph-target (e.g. a 3D model (see Section 3.3)).

4.1 LBP Extraction and KNN Classification

Given an input 2D video performance we first locate the face in each frame. We do this by applying the Viola-Jones face detector [19]. We assume that only one face will appear in each image, and eliminate false-positives by manually identifying the face in the first frame, and then adding the constraint that the correct face in the future frame should be the closest location to the detection in the current frame.

Once we have detected the face we apply Local Binary Patterns (LBP) to the entire image, i.e. across regions 1, 2, 3 and 4. We form these patterns using a pixel's 8 nearest-neighbours. The values generated by the LBPs are then used to create histograms for each facial image, which act as our feature vectors. For a complete description on how to implement LBPs as facial feature vectors the reader is referred to [1]. Note that in this description LBPs are applied to localised block-based regions on the face as opposed to across the entire face. This can be used to retain information relating to the location of facial changes, and can be effective at differentiating between smaller region specific changes.

In our work we restrict automatic recognition to expressions containing distinct global changes defined by combined actions, i.e. we concentrate on recognition of actions 1+2, 3+4, and 5+6, i.e. $SurprisedForehead+Smile$, $SadForehead+Frown$, $AnnoyedForehead + NoseWrinkle$. Therefore, while we acknowledge that calculating LBPs in smaller regions may benefit recognition of localised actions, we have found that for global expression recognition calculating LBPs over the entire face can provide satisfactory results.

For recognition of our combined facial actions we use a K-Nearest Neighbour (KNN) classifier. We assign class labels to our training set of histogram feature vectors depending on what action combination it represents. We do not discriminate between action intensities when applying labels. Therefore, the same label is added to a feature vector representing e.g. a full smile or a partial smile. We describe how we convert a classification result into an expression intensity below.

Given a new facial performance, each input frame is converted in turn into a feature vector and then input into the KNN classifier. Next, instead of strictly assigning the input frame to the class with the majority of the labels, we assign it to a proportion of each class based on all of the K nearest neighbours. We then normalise each proportion by K so as to represent an action's activation level as a value between zero and 1. For example, if $K = 10$ and the K nearest neighbours are all from action pair $1 + 2$, then action $1 + 2$ has a value of 1, and the other actions have zero values. However, if $K = 10$ and 8 of the nearest neighbours are from action $1 + 2$ and the other 2 are $5 + 6$, then action $1 + 2$ has an intensity value of 0.8 and action 5+6 has an intensity value of 0.2.

Viewed as a continuous signal varying over time, the activation level for each

expression can contain noise. This in turn can result in a poor animation. As a post-processing step we therefore median filter each facial action trajectory using a window of size 3, and then fit a polynomial curve to the result. Section 5.1 demonstrates this transition from a noisy to smooth signal in more detail.

Finally, these processed activation levels may then be applied directly to appropriate modes of variation in an appearance model or to morph targets in a 3D model in order to drive an animation.

5 Results

In this Section we first demonstrate animation results. We then examine our approach from a quantitative perspective before considering the performance of our fully automatic performance driven animation technique.

Figure 10 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 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 person's 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 12, and demonstrates 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 a very large proportion of the action's 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 actions 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 [10] where $\mathbf{c}$ is an appearance parameter

$$M(\mathbf{c}_i, \mathbf{c}_j) = \frac{(\mathbf{c}_i \cdot \mathbf{c}_j)^2}{(\mathbf{c}_j \cdot \mathbf{c}_j)(\mathbf{c}_i \cdot \mathbf{c}_i)} \tag{9}$$

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 while not all of the facial actions are completely orthogonal (i.e. having values of zero), many are strongly orthogonal. Specifically, only 6 out of the 45 orthogonality scores have orthogonality values worse than 0.249. This means that during video analysis, many of the actions are be measured from the video with a satisfactory accuracy.

Where two actions have a weak orthogonality, we believe this is caused by them producing similar variations in the same region. Therefore, projecting a face image with this action variation onto the modes of the appearance model

produces substantial weighs on more than one distinct mode as opposed to variation on only 1 distinct mode (which happens in the 100% orthogonal case). However, it should be noted that this does not mean that the action mode will not produce a good representation of the action during animation.

In this discussion of orthogonality, the reader should bear in mind that the modes of variation of our appearance model are always all 100% orthogonal. What we are measuring in this evaluation is how orthogonal our *desired actions* are with respect to these modes. For example, we desire our model to have one distinct mode of variation for a smile and for all other modes to contain no smile variation. However, invariably this is a difficult task, and we would expect at least a small percentage of the smile variation to be present on at least one other mode. It is this action variation which our orthogonality measure is informing us of.

At the beginning of Section 3.1 we described how using local appearance models for facial regions can also provide more useful modes of variation in terms of separating meaningful actions over different facial regions. Figure 11 shows two appearance models constructed using data from only regions 1 and 2. The major actions created in our full-facial model are all present in these modes. In the region 1 model action 1 is present in mode 1 while actions 3 and 5 are present in mode 2. In the region 2 model actions 2 and 4 are present in mode 1 and action 6 is present in mode 2. The missing eyebrow actions, while not present in the region 1 model, could also be simulated by creating appearance models specific to regions 3 and 4.

Upon visual comparison, our new appearance model construction approach compares equally to the local model approach with respect to representing dis-

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

Table 2: Orthogonality between Facial Actions 1 to 10 for our Male Participant. (0 = orthogonal, 1 =non-orthogonal)

tinct region based actions. However, as described in Section 3.1, our new approach combines all the advantages of using local models with the added benefit that only a single full facial model is required as opposed to several different region based models. Additionally, the recombination step required to create the full face from the local models is also not necessary in our new model approach.

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.

5.1 Automatic Performance Driven Animation

We now evaluate the effectiveness of our automatic performance driven animation technique. For this experiment we considered the recognition and transfer of

several different facial actions: $1+2, 3+4$, and $5+6$, i.e. $SurprisedForehead + Smile, SadForehead + Frown, AnnoyedForehead + NoseWrinkle$. We consider these actions in combination since our classifier currently only processes global facial data, and these actions taken separately act on localised facial regions (see Section 3.2 for more details).

We selected approximately 25 training images for each of these three action combinations with which to train our classifier. These image contained representations of these action combinations at several different levels of intensity – ranging from the beginning of an expression to its peak. Given an input test sequence containing similar types of facial behaviour, where each expression was present in durations between 60 and 100 frames, we then automatically classified this in terms of continuous action-combination trajectories over time. In this work, we only consider person dependent classification. Therefore, both training and test images were taken from the same person.

We have briefly experimented with person independent classification by training on one person and then attempting to classify a video sequence of a new person. Test results were mixed. When trained with the male participant, the female $5 + 6$ combination was often misclassified. However, actions $1+2$ were recognised in almost all of our tests. Interestingly, when reversing the test situation (training with female data and testing with male data), the $5+6$ action was classified strongly and the $1+2$ action was classified weakly. We also trained a classifier on both male and female examples. In this test, female $1+2$ expressions were often classified as male $5 + 6$ actions, meaning that identity may be affecting the feature vectors. In summary, these experimental results suggests that LBPs may not be the best features for representing expression

based variation. In this case, features obtained using Gabor filters – as suggested by previous work studying expression recognition – might be better for person independent tasks [2].

We next report our person dependent results. For numerical evaluation we calculated the correlation coefficient between the automatically classified sequence and a ground truth sequence – the latter being generated by semi-automatically tracking the test image set using a facial appearance model (see Section 3). Note that the accuracy of the ground truth sequence still depends on how well our modified appearance model encodes our specific action mode variations. However, the strong level of orthogonality displayed in our models actions satisfies us that even if not 100% accurate, this ground truth is still a good basis on which to perform evaluation and comparison. The alternative would be to hand label each ground truth image with a certain proportion of action variation, which we decided was too difficult.

In order to make this sequence comparable to our automatically classified sequence, we combined the ground truth signals generated by action pairs $1 + 2, 3 + 4$, and $5 + 6$. We did this by simply summing the values of each action signal pair at each time interval and dividing by 2 (giving a value between 0 and 1 for each pair).

We calculated the *correlation coefficient* between automatically generated and ground truth trajectories for each action using

$$C = \frac{\sum_{t=1}^{N}(a_t - \bar{a})(b_t - \bar{b})}{\sigma_a \sigma_b} \tag{10}$$

where N is the length of a trajectory, a_t and b_t are the trajectory values for the

automatic and ground truth sequences at time t, $\bar{a}$ and $\bar{b}$ are the mean trajectory values, and σ_a and σ_b are the standard deviations of each trajectory.

Experimenting with values of K for the KNN classifier ranging from between 4 and 16, we found the best average correlation coefficient was obtained for $K = 10$. This gave us coefficients of $0.96, 0.76$ and 0.83 ($mean = 0.86, std = 0.1073$) for actions $1 + 2, 3 + 4$ and $5 + 6$ respectively. Figure 13 visually compares a set of action trajectories automatically generated via the KNN classifier to a ground truth sequence generated via the modified appearance model.

Our numerical results clearly show a strong positive correlation between the automatically generated sequences and the ground truth – indicating that they are an accurate approximation. Visually, one can also see this correlation. Each action pair curve closely follows its ground truth counterpart. There are however some small deviations (e.g. the beginning of the action 5+6 curve). Improvement of classifier accuracy to overcome these errors is a topic for future work.

6 Conclusions and Future Work

We have presented a method for measuring facial actions from a person's 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 models actions, we have successfully demonstrated that they are well suited to accurately capturing facial actions. This therefore demonstrates the model's suitability for performance driven facial animation applications.

One issue that we currently do not address is the transfer of inner mouth detail e.g. for smiles. This would be an interesting experiment, and it may be the case that further basis vectors 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 [7].

There is plenty of scope to extend our work on automatic facial action detection – the inclusion of local facial actions being an obvious one. This could be approached by constructing feature vectors and classifiers trained especially for localised facial areas [1]. We would also like to extend the size of our training set and also consider other facial actions. Person independent recognition of expressions would also be an exciting challenge.

Acknowledgements

The authors are grateful to Steven Roy for his early technical contributions to this research, and would also like to thank Eva Krumhuber and Gavin Powell for helping us collect the video data used for our model construction.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *Lecture Notes in Computer Science, Proc. of 8th European Conference on Computer Vision*, volume 3021, pages 469–481, 2004.

[2] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognising facial expression: Machine learning and application to spontaneous behaviour. In *Proc. of IEEE Computer Vision and Pattern Recognition*, volume 2, pages 568–573, 2005.

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

[4] Y. Chang and T. Ezzat. Transferable videorealistic speech animation. In *2005 ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, pages 143–152, 2005.

[5] Erika Chuang and Christoph Bregler. Mood swings: expressive speech animation. *ACM Trans. Graph.*, 24(2):331–347, 2005.

[6] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. PAMI*, 23(6):681–684, 2001.

[7] 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.

[8] D. Cosker, S. Paddock, D. Marshall, P. L. Rosin, and S. Rushton. Toward perceptually realistic talking heads: Models, methods, and McGurk. *ACM Trans. Appl. Percept.*, 2(3):270–285, 2005.

[9] P. Ekman and W. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, 1978.

[10] P. Gader and M. Khabou. Automatic feature generation for handwritten digit recognition. *IEEE Trans. PAMI*, 18(12):1256–1261, 1996.

[11] C. Hack and C. J. Taylor. Modelling talking head behaviour. In *Proc. of British Machine Vision Conference (BMVC)*, pages 33–42, 2003.

[12] P. Joshi, 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*, pages 187 – 192, 2003.

[13] F. De la Torre and M. J. Black. Dynamic coupled component analysis. In *Proc. of IEEE Computer Vision and Pattern Recognition*, volume 2, pages 643–650, 2001.

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

[15] I. Matthews and S. Baker. Active appearance models revisited. *Int. J. Comput. Vision*, 60(2):135–164, 2004.

[16] Costen N, T. Cootes, G. Edwards, and C. Taylor. Automatic extraction of the face identity subspace. In *Proc. of British Machine Vision Conference (BMVC)*, volume 1, pages 513–522, 1999.

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

[18] B. Theobald, G. Cawley, J. Glauert, and A. Bangham. 2.5d visual speech synthesis using appearance models. In *Proc. of British Machine Vision Conference (BMVC)*, pages 43–52, 2003.

[19] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

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

[21] L. Williams. Performance driven facial animation. *Computer Graphics*, 24(4):235 – 242, 1990.

[22] 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.

[23] 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.

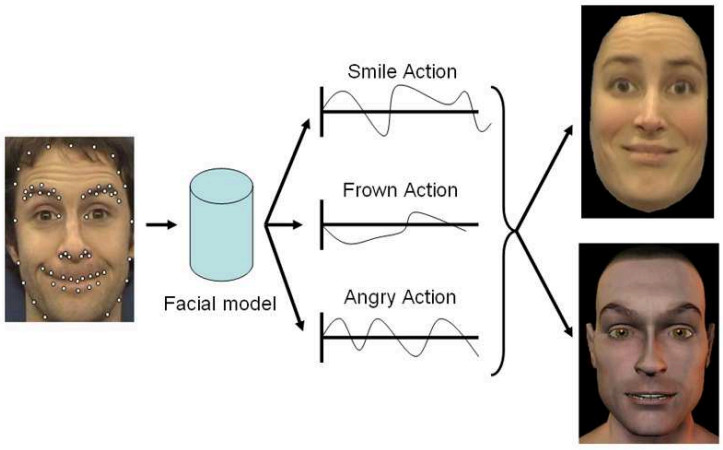

Figure 1: Given a video performance we track facial features and project image and shape information into our facial model. This produces trajectories of facial action parameters which can be used to animate multiple types of facial model, namely appearance models and 3D facial models. Thus, animation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

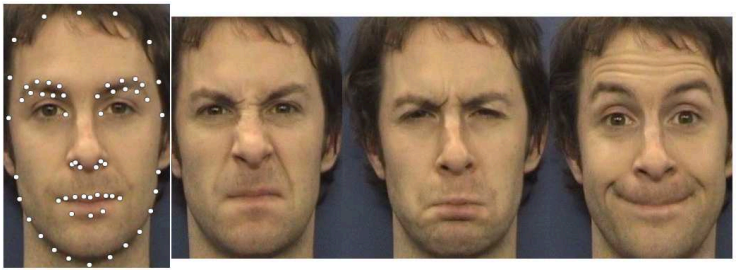

Figure 2: Example landmark placement and participant facial expressions for *Disgust*, *Sadness* and *Happiness*.

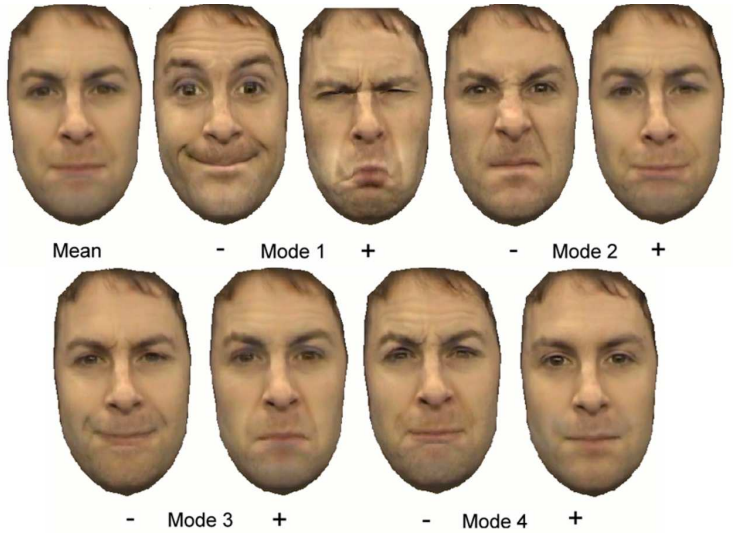

Figure 3: First four modes of variation for our male participant and $\pm$ 3 standard deviations. Note that the modes encode combinations of facial actions.

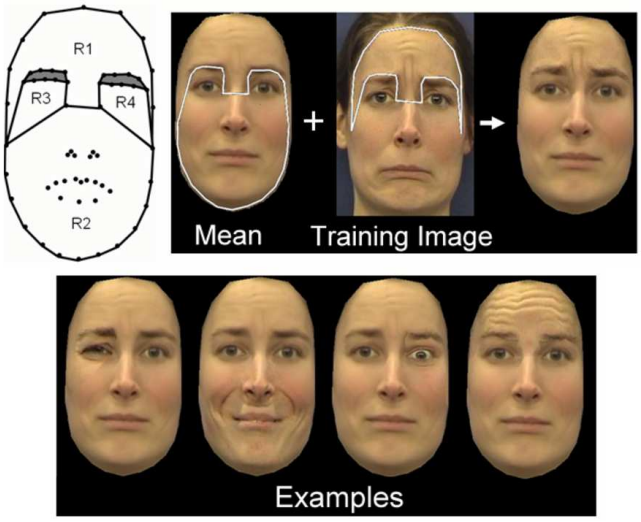

Figure 4: (Left) There are four facial regions. The grey areas are shared by regions R_1, R_3 and R_4. (Middle) An artificial image is constructed by warping a region from a training image over the mean image. (Right) Example artificial images for actions (Left to Right) 8, 2, 9 and 1.

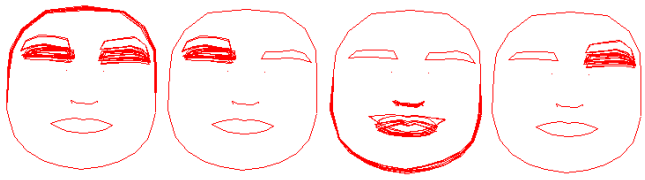

Figure 5: (Left to Right) Superimposed shape vectors for regions R_1, R_3, R_2 and R_4.

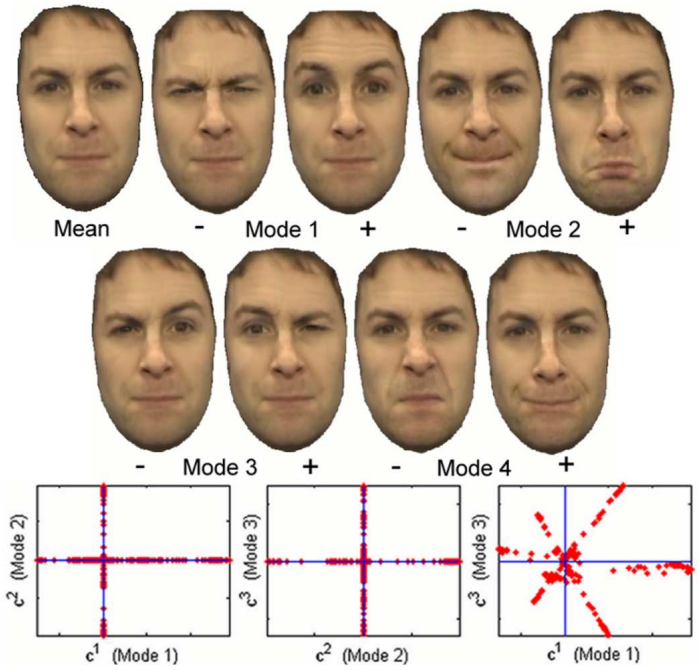

Figure 6: Modes of appearance variation for the new appearance model, along with artificially created training vectors (red-dots) projected into a low-dimensional appearance parameter space. Note how these modes capture localised facial variations, as opposed to the standard appearance model shown in Figure 3.

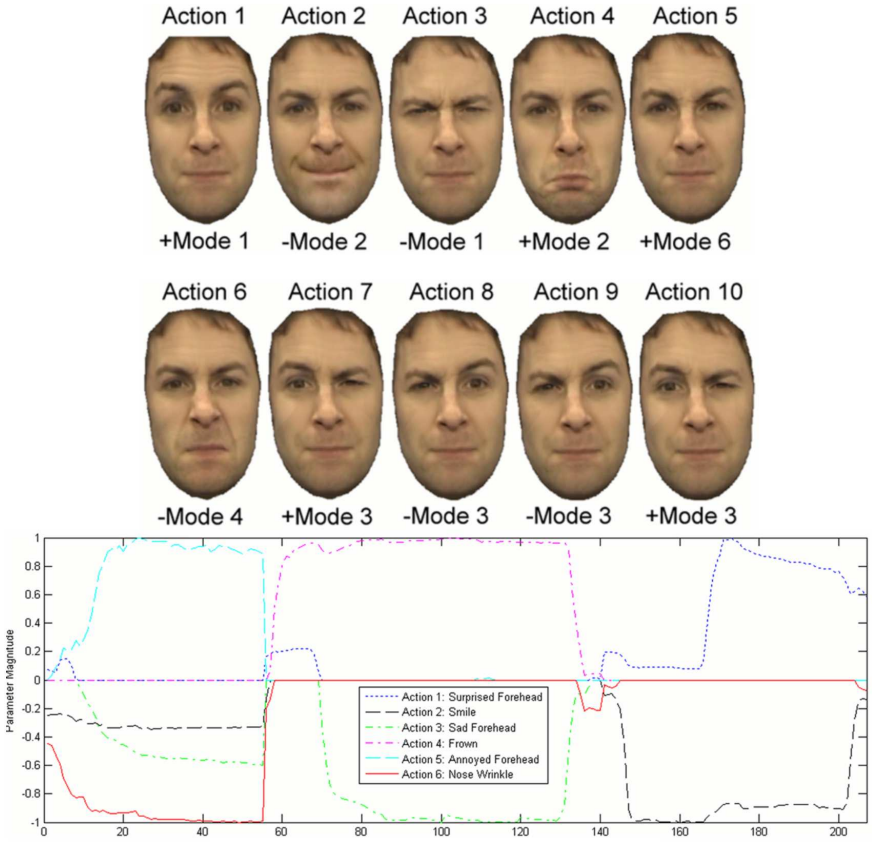

Figure 7: (Top) Modes of variation for the male participant created by our new appearance model. These have been constructed so that they relate to the actions specified in Table 1. (Bottom) A video performance of the male participant represented as a set of continuous action-mode trajectories. The magnitudes of these trajectories are normalised between -1 and 1 using the maximum value for an action-mode (see Section 3.2).

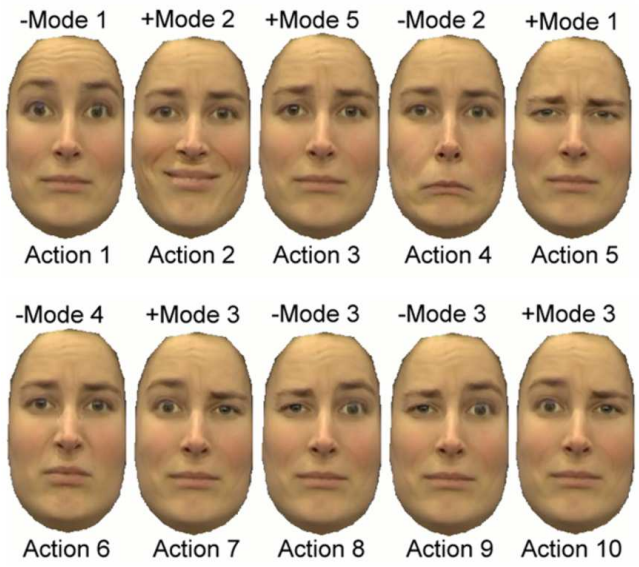

Figure 8: Modes of variation for a female participant created by our new appearance model. These have been constructed so that they relate to the actions specified in Table 1. Animations of this persons face can be created by applying the male action-mode trajectories in Figure 7.

Figure 9: 3D facial model expressions. (Top-row left to right) 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).

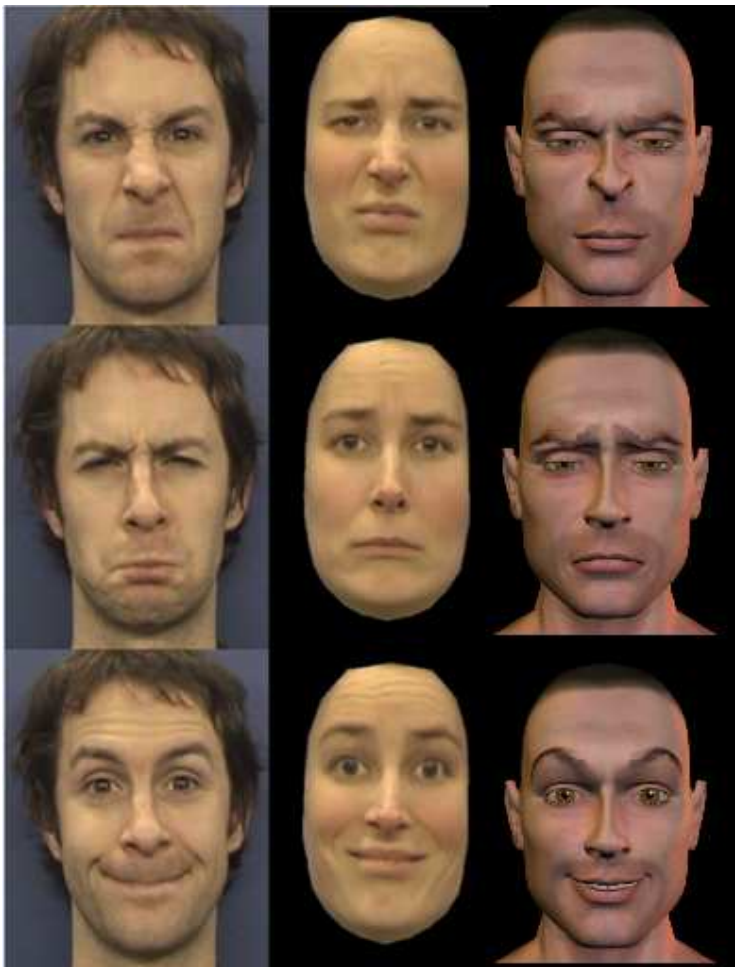

Figure 10: Mapping expressions from our male participant onto both our female participant and a 3D morph-target based facial model.

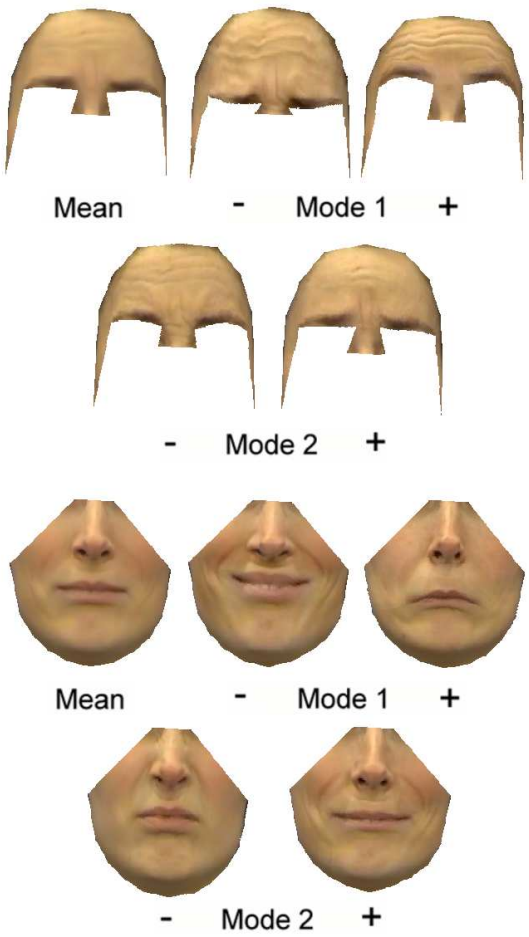

Figure 11: Local appearance models constructed for region 1 (top) and region 2 (bottom). Modes for region 1 are at ± 2.5 s.d while modes for region 2 are at ± 1.5 s.d.

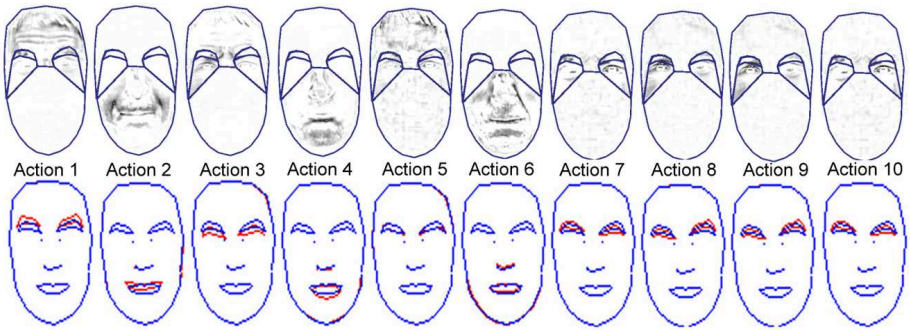

Figure 12: Differences (for the male participant) between texture and shape vectors produced by our actions and their respective means.

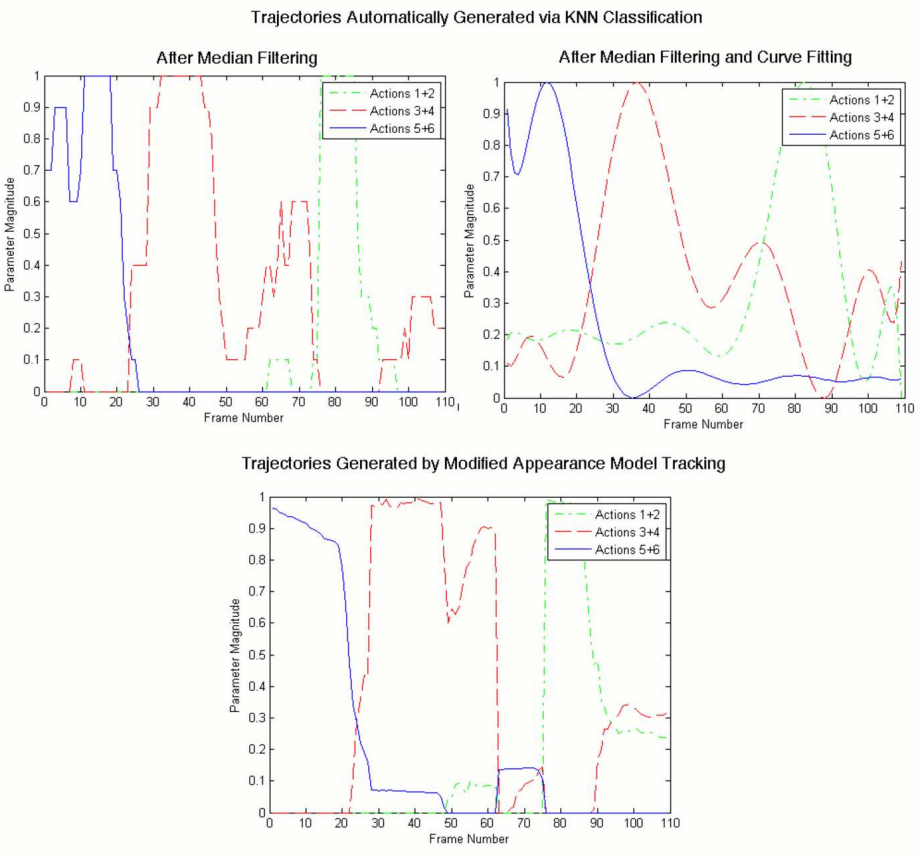

Figure 13: (Top) Visual comparison of action trajectories for KNN classified data after median filtering (left) and after curve fitting (right). (Bottom) Ground truth action trajectories generated by fitting a modified appearance model to the performance.