Synthesis and Recognition of Facial Expressions in Virtual 3D Views

Lukasz Zalewski and Shaogang Gong
Department of Computer Science, Queen Mary, University of London, E1 4NS, UK
[lukas|sgg]@dcs.qmul.ac.uk

Abstract

The human face exhibits complex and rich changes that are both unpredictable and varying in time. In this paper we present a novel method for synthesising and recognition of facial expression changes at extreme 3D views, based on images at near frontal views. Given a sequence of images of facial expressions at near frontal views, we automatically generate virtual expressions at extreme 3D views with corresponding semantic labelling of the expressions. This is accomplished by two components: (1) A shape component where modelling of the shape changes is accomplished through the use of a Mixture of Probabilistic PCA (MPPCA) (2) A texture component where modelling of the semantic changes is performed through auto-clustering of facial expression subspaces in the MPPCA feature space.

1. Introduction

A human face can exhibit complex and intricate expressions. Facial expression changes are dependent on many factors such as muscle contractions, current emotional state and its implied context. Also facial expressions are individually independent: no two people exhibit the same expression in the same way. These factors make modelling and recognising facial expressions a challenging task.

Related to this work Bettinger *et al.*[1] used AAM as the underlying basis of their model, sample mean shift and a variable length Markov model, to learn the relationships between trajectories of facial expressions. Devin and Hogg [6] used AAM combined with sound as their framework to produce sequences of a talking head. Cohen *et al.*[3] used a model based on the motion vectors of Bezier volumes. These vectors were then used in conjunction with a multilevel HMM to classify expressions from image sequences.

Chuang *et al.*[2] used statistical appearance representation (similar to [5]) to represent facial expression configurations, then a factorised bilinear model to synthesise existing sequences with different expressions during the speaking process. Gong *et al.*[8] used Kernel PCA to model nonlinearities introduced by large pose variation. Heap and Hogg [9] introduced a hierarchical combination of linear components in order to model a non-linear manifold.

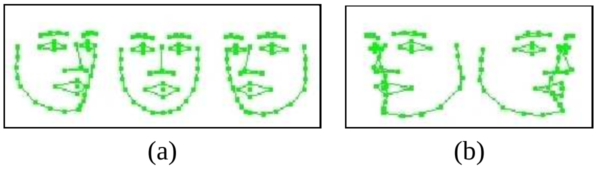
In this work we wish to model and synthesise the appearance and semantics of a set of low-level facial behaviours, including neutral, smile and surprise, across large variations in 3D views. We aim to model the intrinsic inner-expression relationships by placing semantic constraints to bootstrap the process to help in extraction of facial expressions. Facial appearance over varying expressions is based on a statistical appearance model originally introduced by Cootes and Taylor [5].

We extend the basic definition of the model to implicitly incorporate parameters for rotation, scale and large pose variations into the statistical distribution. In order to cope with non-linearities in appearance distribution caused by large pose variation [8], we exploit mixture of PPCA [12]. This approach does not require solving computationally expensive optimisation in reconstruction (as in [8]) and defines a fully probabilistic framework, as opposed to [8, 11]. Another extension of the basic definition of the statistical model is decoupling of the texture model from the shape model (i.e. the model does not control shape and texture variations simultaneously) implying full independence between the two. Facial expressions are grouped into subspaces in texture feature space using MPPCA, and semantic labels of the expressions are then extracted through a Bayesian framework.

2. Our Model

We adopt the Active Appearance Model [5] as the base representation, but extend our model to incorporate expression changes under large pose and scale variations. As Gong *et al.*[8] point out, large pose variations cause the shape space to become highly non-linear. Hence the linear mapping used to model the manifold is no longer sufficient, as illustrated in Figure 1. On the other hand, shape distribution forms distinctive bands of points in the PCA space with respect to the pose. Figure 2 shows projections of the data onto the first three principal axes. The grouping was performed with respect to Y-axis rotation. Triangles represent $[-40^o, -20^o]$, crosses $[-10^o, 10^o]$ and circles $[20^o, 40^o]$ ranges respectively.

Based on this observation we employ a mixture model capable of capturing non-linear distributions in a unified



(a) (b)

Figure 1: Non-valid shape reconstructions of profile views (b) These were generated using modes of variations of a linear shape model trained at near frontal views (a).

framework. We model the shape space using a mixture of PPCA [12]. Cootes and Taylor [4] used Gaussian Mixture Model to define plausible shape space, but our main aim is to segment the feature space according to pre-defined semantics (the definition of plausible shape space is a by-product of such segmentation). The dimensionality of each of the shape mixture components account for 98% of the variation within the training set, whilst the number of mixture components is determined by the number of rotational bands shown in Figure 2 (triangles, crosses, circles).

We make an assumption that texture is independent of shape. This is because we believe that all the necessary information concerning expressions caused by muscle motion can be more effectively modelled by texture values alone, provided that shape variations are normalised. Our motivation for such a decomposition is as follows: First of all the dimensionality of our shape data can be significantly reduced by taking the smallest number of points needed to describe the pose variation and allow efficient warping. Therefore the number of principal components needed to approximate all necessary shape variations is reduced. Secondly, decoupling shape and texture allows us to warp any texture to any possible shape, and decreases the number of shape-texture pairs needed to represent a facial expression across the whole pose sphere. Thirdly, the separation creates two independent sets of expressions: the ones that are derived solely from texture variations, and the ones that are implied solely by pose changes.

In the texture model, each of the shape-free texture vectors is segmented into three regions, which correspond to upper, mid and lower part of the face. The texture feature space corresponding to each of the face segments is clustered according to the expression groups to be modelled, such that each of the clusters represents data distribution for a particular expression. Figure 3 shows shape-free texture vectors of the lower segment of the face projected onto the first three principal components. Crosses correspond to the surprised state, circles to the smile and triangles to the neutral state. Figure 4 shows the modes of variation (± 2.5 standard deviation) for the lower face segment for neutral, smile and surprised expressions (top to bottom). A sequence of facial expression changes can be represented as a path or trajectory in texture feature space, travelling through different expression subspaces. In contrast to [1] we use texture parameter space to obtain the necessary expression labelling.

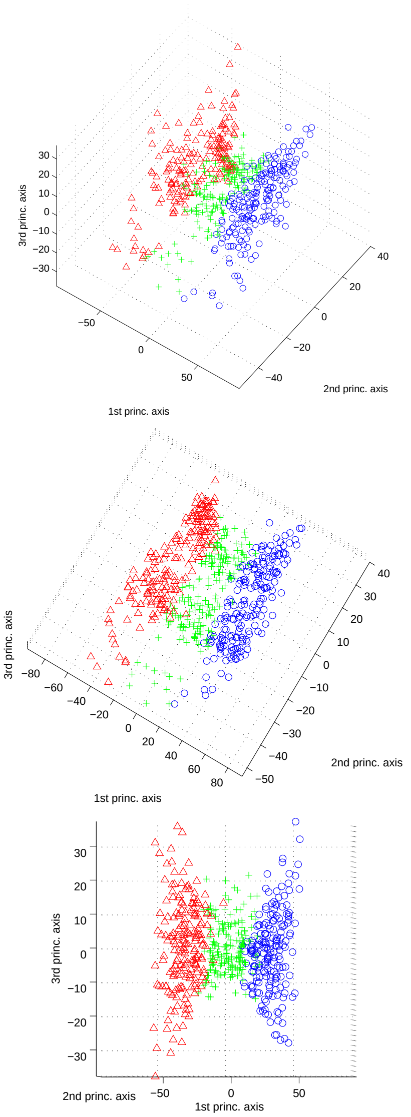


Figure 2: The shape variation of facial expression images from $[-40^o, 40^o]$ 3D views (in yaw) projected onto the 1st three principal components. The manifold forms continuous and separable clusters: $[-40^o, -20^o]$ (shown by triangles), $[-10^o, 10^o]$ (shown by crosses) and $[20^o, 40^o]$ (shown by circles)

The data distribution for both shape and texture is modelled using Mixture of Probabilistic PCA (PPCA is described in Section 2.1). Experiments are presented in section 3 before conclusions are drawn in Section 4.

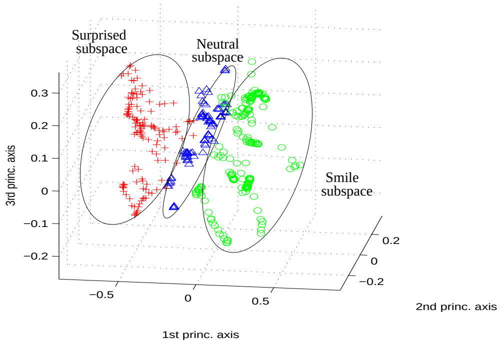


Figure 3: The projection of shape-free texture vectors of the lower face segment form separable subspaces corresponding to different expressions. Three separable expression subspaces (surprised, neutral, smile) are shown by clusters of crosses, circles and triangles respectively.

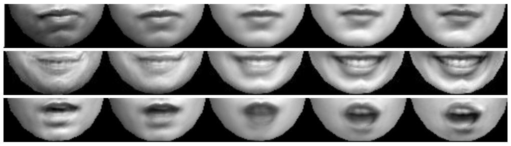


Figure 4: Modes of variation of the bottom face segment (± 2.5 standard deviation) for the neutral, smile and surprised clusters (top to bottom).

2.1. Probabilistic PCA

Due to its linear nature, PCA performs poorly in modelling of the non-linear manifolds, its lack of probability distribution makes it ill-suited for a Bayesian framework. Tipping and Bishop [12] reformulated PCA as the maximum likelihood solution using a latent variable model such that the observed variable $\mathbf{t}$ is given by:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{x}$ is the latent variable such that $P(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}_q)$ and $\mathcal{N}$ denotes a Gaussian distribution, $\mathbf{W}$ is the parameter matrix whose columns define principal subspace of the data, $\boldsymbol{\mu}$ is the d-dimensional vector, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$ where σ^2 is the noise variance, $\mathbf{I}$ is the identity matrix and $\mathcal{N}$ represents Gaussian distribution. Then

$$P(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}_d) \tag{2}$$

Marginal distribution of the observed variable $\mathbf{t}$ is

$$P(\mathbf{t}) = \int P(\mathbf{t}|\mathbf{x})P(\mathbf{x})d\mathbf{x} = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \tag{3}$$

where covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}_d$. The above model represents a constrained Gaussian distribution controlled by $\boldsymbol{\mu}, \mathbf{W}$ and σ^2. A maximum likelihood solution for the parameters is given by:

$$\begin{aligned} \boldsymbol{\mu}_{ML} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{t}_i && (4) \\ \mathbf{W}_{ML} &= \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I}_q)^{\frac{1}{2}}\mathbf{R} && (5) \\ \sigma^2_{ML} &= \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i && (6) \end{aligned}$$

where $\mathbf{t}_i$ is the i-th d-dimensional feature vector from the data set, $\boldsymbol{\Lambda}_q$ is a diagonal matrix containing the q largest eigenvalues λ_i, $\mathbf{U}_q$ is the matrix containing the q largest eigenvectors and $\mathbf{R}$ is an arbitrary orthogonal rotation matrix. Thus the mixture model $p(\mathbf{x})$ can be defined as follows:

$$p(\mathbf{x}) = \sum_{i=1}^{K}\pi_i p(\mathbf{t}_n|i) \tag{7}$$

where $p(\mathbf{t}_n|i)$ is the single PPCA model and π_i is the corresponding mixing proportion, $\pi_i \geq 0$, $\sum \pi_i = 1$.

2.2. Labelling Facial Expressions

Given a shape-free texture vector $\mathbf{t}^{(i)}$ belonging to the i-th face segment, the expression within segment i can be classified by determining its association with a particular cluster by calculating the posterior probability as follows:

$$P(j|\mathbf{t}^{(i)}) = \frac{P(\mathbf{t}^{(i)}|j)P(j)}{P(\mathbf{t}^{(i)})} \tag{8}$$

Then the posterior probability values for all the clusters corresponding to each of the face segments for a given image form a probability matrix $\mathbf{Z}$ such that

$$\mathbf{Z} = \begin{bmatrix} p^{(eye)}_{neutral} & p^{(eye)}_{surprised} & p^{(eye)}_{smile} \\ p^{(nose)}_{neutral} & p^{(nose)}_{surprised} & p^{(nose)}_{smile} \\ p^{(mouth)}_{neutral} & p^{(mouth)}_{surprised} & p^{(mouth)}_{smile} \end{bmatrix} \tag{9}$$

For classification we define a probability weight matrix, in which values are deterministically set based on the amount of contribution of each of the face segments toward the specific expression:

$$\mathbf{W}_p = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0.3 & 0.0 & 0.7 \\ 0.0 & 0.2 & 0.8 \end{bmatrix} \tag{10}$$

The final classification is performed according to:

$$\underset{1 \leq i \leq 3}{argmax} \; \mathbf{L}_t(i,i) \; , where \; \mathbf{L}_t = \mathbf{W}_p\mathbf{Z} \tag{11}$$

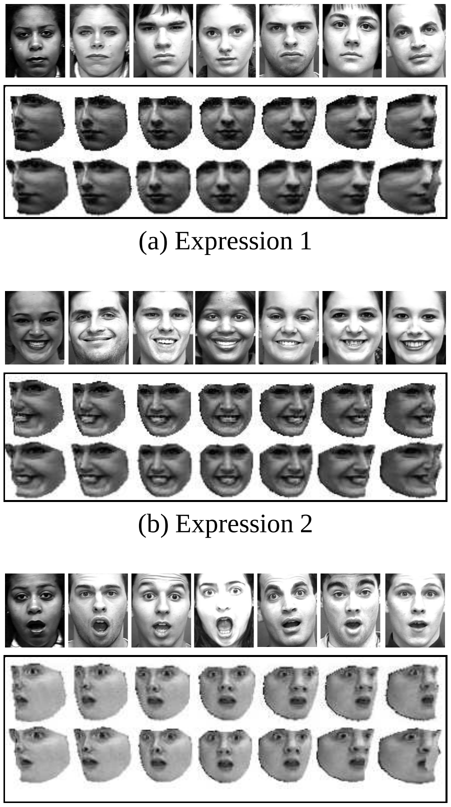
3. Experimental Results

Our shape training set consists of 576 shape samples at near frontal views which cover the range $\pm 40^o$ yaw and $\pm 20^o$ pitch. We employed a PPCA mixture model to capture their manifold distributions in feature space. The resulting dimensionality of the components were set to 10 (98% variation of the training set). It yielded a single model covering a large view sphere in a unified probabilistic manner.

Figure 5 shows examples of synthesising three different types of facial expressions, from near frontal views to profile views (± 2.5 standard deviation). The top row shows examples of different expressions from the training set covering $\pm 10^o$ views. The middle row corresponds to morphing performed with the MPPCA model, whilst the bottom row shows the morphing with the PCA based model with visible kinks due to inability of the model to cope with nonlinear shape space. Figure 6 shows examples of morphing to extreme virtual 3D views (± 3.0 standard deviation) using different texture vectors. In each column of (a) and (b), images on the left were generated using the PCA model (distortions present), and the images on the right using a MPPCA model.

The expression set consist of 490 images (training set) and 300 images (testing set) from the Cohn-Kanade Facial Expression Database [10]. Our shape-free texture vectors are obtained by morphing all the texture vectors onto the mean shape (details about morphing methods can be found in [5]). Once morphed, we divide each of the patches into three segments corresponding to the upper, mid and lower part of the face. Since the facial motion of the lower part of the face has little influence over the motion of the upper part [7], we impose three-part decomposition which aims to introduce semantic correlations between upper-middle and lower-middle parts of the face, and at the same time to reduce misclassification caused by visual ambiguities. Additionally, such segmentation reduces overall dimensionality of the space and the number of sample combinations needed to describe a particular expression.

During our experiments we noticed that for particular facial expressions, only a few out of all the segments convey relevant information, and the remaining ones can be discarded. For example when we smile, only lower and mid segments can be used for classification purposes (mouth shape and possible skin creases around the nose area), and when we are surprised, relevant information is mostly conveyed through mouth shape and widening of the eyes. Figure 7 shows different motion areas for different types of expressions. We can see that for the smile expression (a) the motion is mostly concentrated around the mouth and nose areas, and for the surprised expression (b) concentration falls into the mouth and eyes region.



(a) Expression 1

(b) Expression 2

(c) Expression 3

Figure 5: Examples of morphing (synthesising) facial expressions into extreme virtual views. The top rows in (a), (b), (c) are some of the training examples of three different expressions at near frontal views. The middle rows were computed using MPPCA model. The bottom rows were computed using the PCA model with visible kinks at extreme 3D views (profile views) due to non-linearities present.

Each of the face segments was modelled using MPPCA with a number of components corresponding to the number of facial expression states we wish to model (neutral, smile, surprise). Figure 8 shows two probability plots generated by the texture model (Equation (11)) from two novel testing sequences exhibiting continuous changes of expression from neutral to smiling (bottom plot) and from neutral to being surprised (top plot). In each of the plots the solid line represents the probability of the expression being classified as neutral, dashed line as surprised and dash-dot line as smiling. Each of the images within each of the plots shows the expression synthesised at a virtual 3D view (image on the left), the original image (middle) and three part decomposition of the texture vector with posterior probability values (Equation (8)) for the classified expression.

Figure 9 shows the results from those two test sequences projected into the feature space of the lower face part, with the left plot corresponding to the individual being sur-

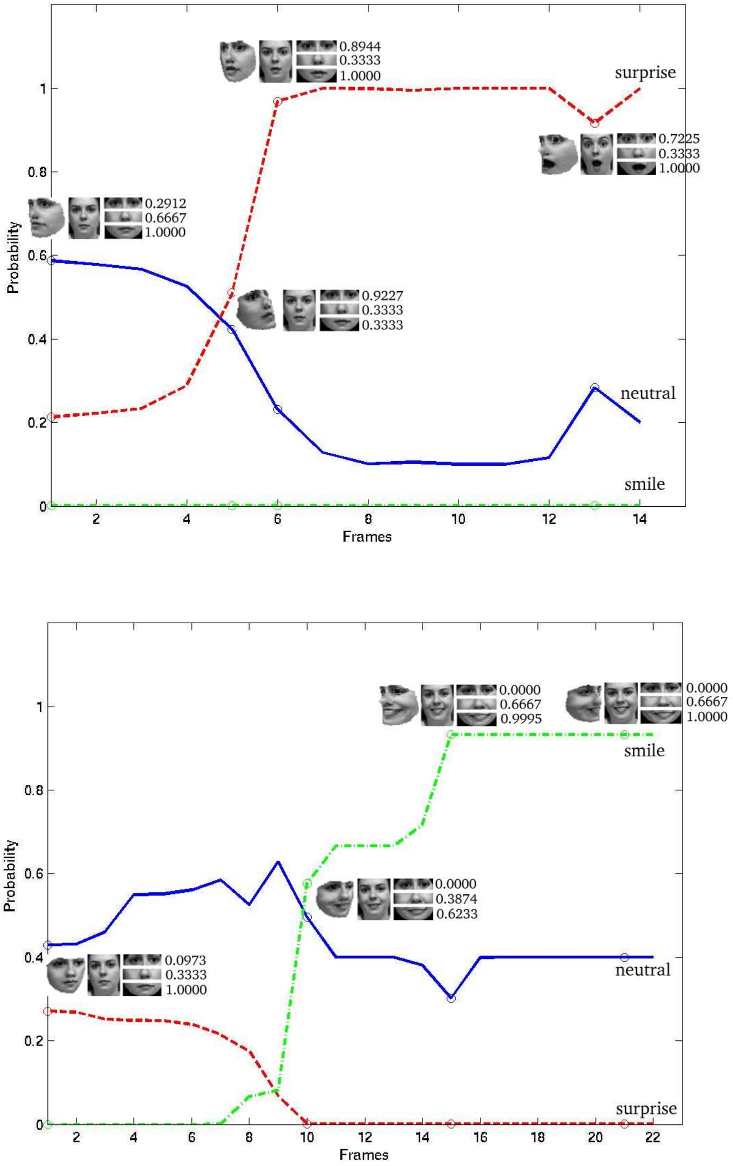


Figure 8: Expression recognition probability values estimated for two example test sequences: surprised (top plot) and smile (bottom plot). The solid line shows the probability of the expression being classified as neutral, dashed line as surprise and dash-dot line as smile. Each of the images within the plots shows the synthesised expression at a virtual 3D view (on the left), original image (middle) and three part decomposition (right) of the texture vector, with posterior probability for the currently classified expression shown for each of the segments.

prised, and the right one to smiling. It can be seen that the sequences form visible trajectories (solid line) travelling through different expression subspaces.

4. Conclusions and Future Work

In this paper, we have shown a general probabilistic framework for synthesising the shape and texture variations of facial expressions from near frontal views ($\pm 10^o$) to extreme virtual views. We demonstrated the advantages of using MPPCA over PCA for this task. We have shown that the shape and texture can be treated as independent entities, and modelled as such, and that facial expression synthesis can be accomplished by using shape as a basis to morph the texture onto, which in turn can be obtained by traversing texture parameter space. We also showed that classification

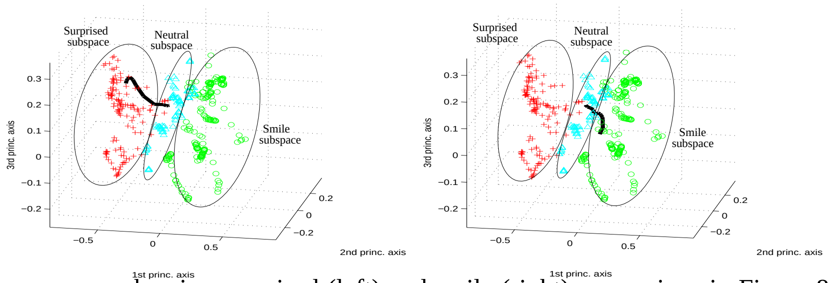
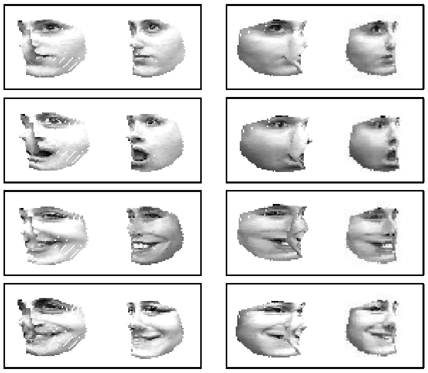


Figure 9: The two test sequences showing surprised (left) and smile (right) expressions in Figure 8 projected onto feature space of the lower face part, and showing the trajectories of expressions travelling through different expression subspaces.



(a) (b)

Figure 6: Examples of synthesising the texture vectors of different expressions to extreme virtual 3D views (± 3.0 standard deviation from the trained model). Column (a) represents -3.0 standard deviation, column (b) represents $+3.0$ standard deviation. In each column images on the left were generated using a PCA model, images on the right using a MPPCA model.

of the expressions can be performed in a probabilistic manner using a unified probability distribution function. Our future work includes use of a larger set of facial expressions, and utilisation of temporal information as an extension of the current model to define the dynamics of the expressions.

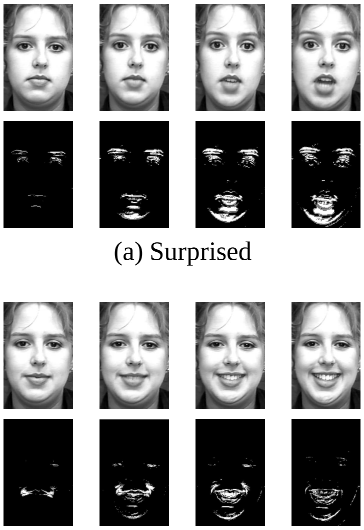
References

[1] F. Bettinger, T. F. Cootes, and C. J. Taylor. Modelling facial behaviours. In *BMVC*, volume 2, pages 797–806, 2002.

[2] E. S. Chuang, H. Deshpande, and C. Bregler. Facial expression space learning. In *10th Pacific Conference on Computer Graphics and Applications*, Beijing, 2002.

[3] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences. *International conference on Multimedia and Expo*, 2:121–124, 2002.

[4] T. Cootes and C. Taylor. A mixture model for representing shape variation. In *BMVC*, volume 1, pages 110–119, 1997.



(a) Surprised

(b) Smile

Figure 7: Original image (top rows in (a) and (b)), and the regions of the image exhibiting motion during expressions (bottom rows) such as surprised (a) and smile (b).

[5] T. F. Cootes and C. J. Taylor. Statistical models of apperance for computer vision. Technical report, University of Manchester, Manchester, UK, 2001.

[6] V. E. Devin and D. C. Hogg. Reactive memories: An interactive talking head. In *BMVC*, 2001.

[7] G. Donato, M. S. Barlet, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *PAMI*, 21(10):974–989, October 1999.

[8] S. Gong, A. Psarrou, and S. Romdhani. Corresponding dynamic appearances. *Image and Vision Computing*, 20:307–318, 2002.

[9] A. Heap and D. Hogg. Improving specificity in PDMs using a hierarchical approach. *BMVC*, 1:80–89, 1997.

[10] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis, 2000.

[11] Y. Li, S. Gong, and H. Liddell. Constructing facial identity surfaces for recognition. *IJCV*, 53(1):71–92, 2003.

[12] M. E. Tipping and C. M. Bishop. Mixture of probabilistic component analysers. Technical report, 1998.