Generating a Mapping Function from one Expression to another using a Statistical Model of Facial Shape

John Ghent
Computer Science Department
National University of Ireland, Maynooth
Maynooth, Co. Kildare, Ireland
jghent@cs.may.ie

John McDonald
Computer Science Department
National University of Ireland, Maynooth
Maynooth, Co. Kildare, Ireland
johnmcd@cs.may.ie

Abstract

We demonstrate a novel method of generating a mapping function from the shape of a neutral face to the shape of one showing an expression. It is proposed that this mapping function can be used to automatically generate facial expressions from still images of never seen before faces or classify which expression a person is portraying. This new technique draws on the work of Ekmans [8] *Facial Action Coding System* (FACS), Cootes [4,5] *Active Shape Model* (ASM) and Artificial Neural Networks (ANN). To build an ASM it is required to have a training phase, where each image in the training phase is 'scored' by the FACS system and a neural network is used to generate a mapping function as a face moves from a neutral expression to an alternative expression. We describe this method in detail and give results indicate the effectiveness of the technique.

Keywords: Facial Expression Generation, Active Shape Model (ASM), Facial Action Coding System (FACS), Principle Component Analysis (PCA), Feedforward Heteroassociative Memory Network.

1 Introduction

The purposes of this paper are, firstly to demonstrate how facial expressions can be modelled in a consistent manner, secondly, to build a statistical model of facial expressions, and thirdly, to show that a mapping may be derived that maps an image corresponding to a neutral expression to an image corresponding to a non-neutral expression. We address the problem by generating a mapping from a neutral expression to a non-neutral expression independent of the subject. This is achieved by building a statistical model of the expression in question from a number of subjects showing that expression in a training set. The change in shape of each face in the training phase is then analysed and used to derive a mapping function, which takes their neutral face to one depicting the new expression. This mapping function from one expression to another is a difficult problem. To decrease the dimensionality of this mapping the variance in shape of each face in the training set is analysed using principle components analysis [10]. This approach can model a large amount of the variance in the training set by using only a few principle axes or principle components. In other words, the goal is to obtain a smaller set of features or vectors that reduces the complexity of the mapping but can accurately represent the original training set.

Before a model of facial expression can be developed, a system measuring facial expressions would have to be in place. This is necessary for measuring the accuracy of results and for allowing consistency in expression description regardless of identity. This is achieved using the *Facial Action Coding System* (FACS) developed by Ekman [8]. The FACS system separates the muscles in the face into Action Units (AU) and uses these AU's as a basis for facial movement.

As with many vision and graphics applications, the generation of facial expression is often a difficult computational problem. Solutions to this problem are often are too complicated to run in real time and therefore cannot be used in adaptive applications. Those that can run in real time are often inaccurate and could never be used for applications that require a high level of robustness. Picard [20] details a procedure that has a 98 percent accuracy rate at classifying facial expressions but takes five minutes to process a facial expression. A method has been developed that augmented Ekmans FACS for facial expression and recognition using templates of motion energy, but this cannot be run in real time. Yaser Yacoob and Larry Davis of the University of MaryLand, use templates of motion energy, but use a combination of templates and smaller sub-templates and combine them with rules to formulate expressions [20]. This method is also not real-time, because computing the motion flow is very time consuming.

The question now becomes: what is the most effective, efficient and accurate method or combination of methods suitable for automatically generating and recognizing arbitrary facial expressions in real time? A method is proposed in this paper that details a procedure that provides accurate results that could be run in real time.

This technique could be used in many applications such as an arbitrary animated agent (a photo realistic animation of facial expression), automating an interactive web host (instead of streaming video across a network, a single image could be transmitted and the image could be manipulated on the machine the user is using), face and expression verification (explained below), and adding personality and emotional expressions to an arbitrary image of a face.

Other useful applications would be in the field of verification and authentication. This paper shows that there is a distinct change of facial parameters regardless of identity (see section five). This suggests that any expression can be automatically classified or generated using this novel technique. This has useful implications, if a user of a system had a web camera mounted on top of their monitor, this technique can be used to correctly identify that persons expression. It is important to note that this is not facial recognition but expression recognition. However, if this mapping could be inverted, the technique could be could be used in facial recognition systems.

2 Measuring facial expression

An understanding of how the face creates expression is necessary before a model of facial expression can be generated. As everyone's facial features are unique, the only feasible way to measure an expression is by the movement of the muscles on the face. For this reason the anatomy of the face is a very important aspect in understanding expression

2.1 Facial Action Coding System (FACS)

The FACS is based on an anatomical analysis of facial actions [8]. A movement of one or more muscles of the face is known as an action unit (AU). Sometimes it is hard to distinguish if a set of muscles are accountable for an expression or if a single muscle is, for this reason the term 'Action Unit' is used. All resulting expressions can be described using the AU's or a combination of the action units. The FACS System can account for over 10,000 expressions.

3 Building a statistical model

In order to build a reliable model of facial expression, the model must provide adequate flexibility to cover the vast differences in human expressions. However, it was also important that the models only deform in ways consistent with the FACS. In other words the model should only be allowed generate expressions that are consistent with the training set. The following sections details how such a model is built.

3.1 Labelling the training set

In order to represent a shape we first have to label it with a set of points. These are located around the face and around key areas such as the eye, nose, mouth and eyebrow. These points are weighted on their level of importance, depending on which AU the statistical model is being built for. For example, if a model were being built for AU 23 (Orbicularis Oris) then the points around the mouth would have a greater weight than the points around the eye. There are 122 points in total, 27 around the outline of the face, 24 around the mouth, 15 around the nose, 8 around each eye, 8 around each pupil and 12 around each eyebrow. Fig 1 shows a labelled face.

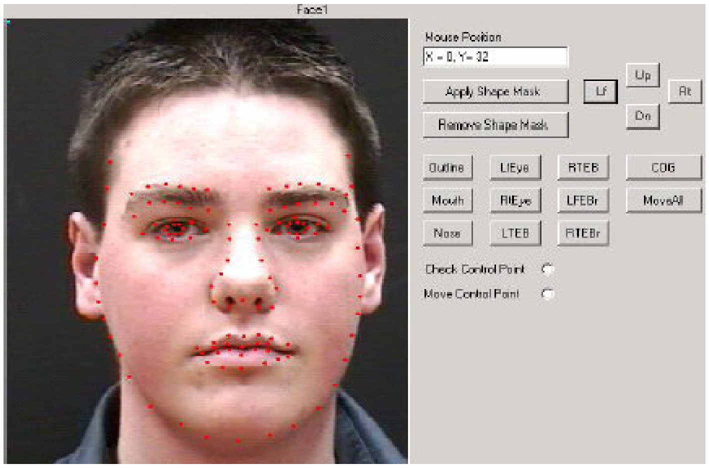
Mouse Position
X = 0, Y= 32

Fig 1. A Screen shot of the software used to label the face

3.2 Aligning the training set

To analyse the variance of the points that describes the shape of the face it is necessary that the faces in the training set are as closely aligned as possible. One way of doing this is to use a technique called the generalized procrustes alignment (GPA)[14]. This technique aligns two shapes with respect to position, rotation and scale by minimizing the sum of the squared distances between the corresponding landmark points discussed in the previous section. The alignment will depend on the weights given to each of the points, which in turn depends on which synthetic AU is being generated. To align two shapes $\mathbf{x}_1$ and $\mathbf{x}_2$, we choose a rotation, scale and translation that minimises the sum of the squared distances between $\mathbf{x}_1$ and $\mathbf{x}_2$. Let $\mathbf{x}_i$ be

$$\mathbf{x}_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, \cdots, x_{in-1}, y_{in-1})^T \tag{1}$$

where *n* is the number of points used to describe the shape of the face. In this case *n* is 122.

Equation (2), is the rotation, scale and translation equation that is to be minimised [3].

$$E = (\mathbf{x}_1 - M(s,\theta)[\mathbf{x}_2] - t)^T \mathbf{W} (\mathbf{x}_1 - M(s,\theta)[\mathbf{x}_2] - t) \tag{2}$$

$M(s,\theta)[\mathbf{x}]$ is the rotation θ, and scale *s* of **x**, an *t* is a translation of **x**.

The complexity of the mapping to be learned depends on the way it is represented. In this case the mapping of an expression change needs to be independent of identity. For each identity, the neutral expression is aligned to the non-neutral expression using the labelled points that do not move as an expression is formed. For example, if AU 4 (brow lowerer) were being mapped, the neutral face and the face showing AU4 would be aligned using every point except the points along the eyebrow. This will emphasize the variance along the eyebrow as that expression is formed and allow for a less complex mapping function to be generated. This realignment is the key aspect in finding a mapping function for this expression that is independent of identity.

3.3 Principle Component Analysis (PCA)

Before any significant analysis can be performed on the shape of the faces, the mean must be acquired. This is found using

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{3}$$

where N is the number of images in the training set.

The way in which the landmark points tend to move with respect to each other can be analysed using a method known as *Principal Component Analysis* (PCA, also known as the Karhunen-Loéve transform). This method takes a set of data points and constructs a lower dimensional linear subspace that bests describes the variation of these data points from their mean. This is computed by

$$\partial\mathbf{x}_i = \mathbf{x}_i - \overline{\mathbf{x}} \tag{4}$$

where (4) represents the difference the first face has with is corresponding points in the mean. A *2n x 2n* covariance matrix can be calculated using:

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\partial\mathbf{x}_i\partial\mathbf{x}_i^T \tag{5}$$

The eigenvalues and eigenvectors of the covariance matrix are then calculated. The eigenvector corresponding to the largest eigenvalue describes the most significant mode of variation. Only a few principle axes are needed to describe the majority of variance of the training set. If **P** is the set of eigenvectors and **b** is a vector of weights, then new shapes can be generated using equation (6):

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{Pb} \tag{6}$$

where **P** is a *2n* x *m* matrix, **b** is a *m* x *1* vector and *m* << *2n*.

4 Generating a mapping

To compute the mapping we use a *Feedforward Heteroassociative Memory Network* (FHMN) [19]. This kind of network is simply a one-layer network that stores patterns. FHMN include the class of networks known as content addressable memories or memory devices that permit the retrieval of data from pattern keys that are based on attributes of the stored data. A simple FHMN is shown in Fig 2.

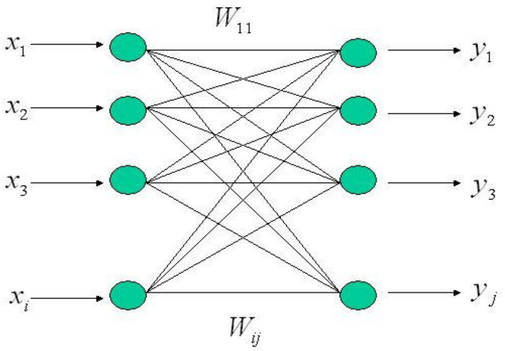

Fig 2. A Feedforward Heteroassociative Network

As the diagram suggests, if the input to a group of neurons is **x** and the output excitations are **y** = **y(x),** then the synaptic weight values W_{ij} are given by

$$W_{ij} = \alpha x_i y_i \tag{7}$$

where α is a normalizing constant.

To retrieve a pattern $\mathbf{y}^k$, the associated pattern $\mathbf{x}^k$ is input to the network, thus:

$$\mathbf{W}\mathbf{x}^k = \left(\sum_{P=1}^{P} \mathbf{y}^P \left(\mathbf{x}^P \right)^T \right) \cdot \mathbf{x}^k \tag{8}$$

In other words, if we had five input-to-output associations to be learned, we would first compute each $\mathbf{W}^i$ by using $\mathbf{W}^i = \mathbf{y}^i (\mathbf{x}^i)^T$, then create the final matrix **W** using $\mathbf{W} = \mathbf{W}^1 + \mathbf{W}^2 + \cdots + \mathbf{W}^5$. It is then possible to retrieve patterns using equation (8).

5 Experiments and results

Both of the experiments described in this paper demonstrate two distinct properties of this technique. The first experiment demonstrates that using a small training set and a subtle change in facial expression it is still possible to generate a mapping function. The second experiment demonstrates how a mapping can be generated for a more substantial expression change such as a smile and shows how this can be used to predict the most probable change in facial shape in never seen before facial images.

5.1 Experiment one

Ten people were used to create a mapping from a neutral face to one showing AU 4. Each image was taken using a Sony EVI-D31 CCD camera. The lighting, pose and orientation of the faces were kept consistent so that the variance measured would be due to change in expression and identity only. Each person was asked to learn AU 4 as described by the FACS system and perform the action unit with intensity *C*. AU4 is a ‘brow lowerer’. There are 5 intensities *A,B,C,D,E*, where *A* stands for the minimum, up to *E,* the maximum intensity. AU *C* is the most common so this is what was modelled.

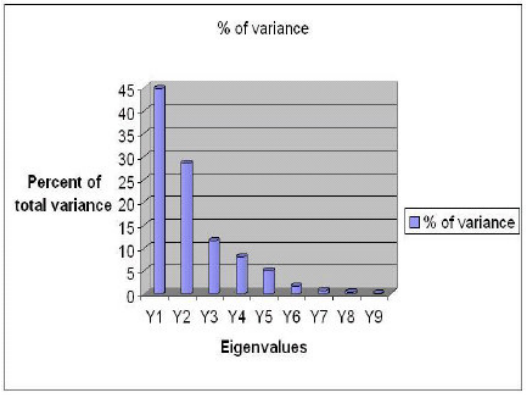

Fig 3. Percentage of eigenvalues

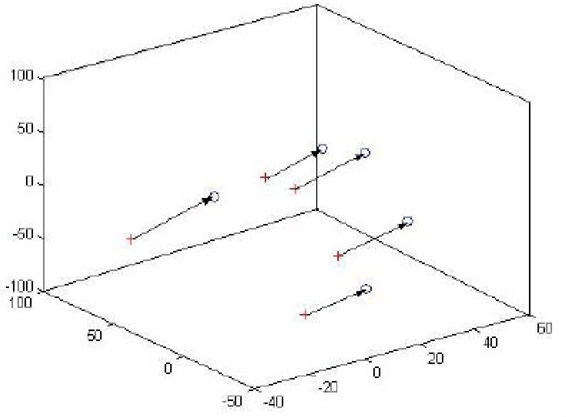

Fig 4. Mapping taken from AU0 to AU4 in eigenspace

Each face was labelled using 122 points, as in Fig 1. They were aligned with each other using procrustes alignment and then for each identity the neutral expression was aligned to the non-neutral expression in a way that best showed the variance of shape across the eyebrows. PCA was performed on the data and the top three eigenvalues were used to generate the appropriate mapping. The top three eigenvalues explain 85.05% of the total variance, as shown in Fig3.

A mapping was then generated using the top three eigenvalues, which represented the neutral faces as input, and the largest three eigenvalues that represented the faces showing AU4*C* (action unit 4 with intensity *C*) as output. A FHMN was used to generate the mapping. From equation (6), the parameters **b**, used for the inputs and outputs can be acquired using

$$\mathbf{b} = \mathbf{V}^T\left(\mathbf{y} - \overline{\mathbf{x}}\right) \tag{9}$$

where **V** are the eigenvectors which explain the variance shown in the training set and **y** is the shape of a face. Fig 4 shows the mapping for a number of subjects from a neutral expression to AU4. The '+' sign represents the input vector (the largest three eigenvalues used to portray AU0 or a neutral face for a specific identity) and the 'o' symbol represents the output vector (the largest three eigenvalues used to portray AU4 for a specific identity).

Once the mapping was created the model was tested by putting the inputs through the mapping function and comparing them to the original outputs. The output **b** of each input was converted into a facial shape using equation (6). Fig 5 shows the results of putting the original input vectors through the mapping function.

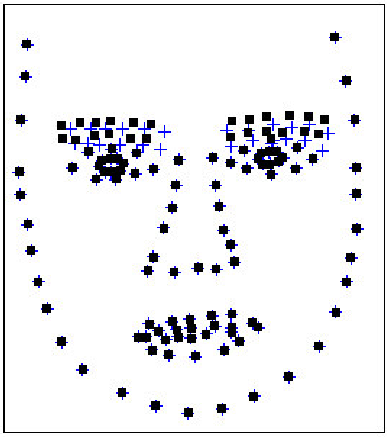

Fig 5.AU0 image superimposed over AU4 image

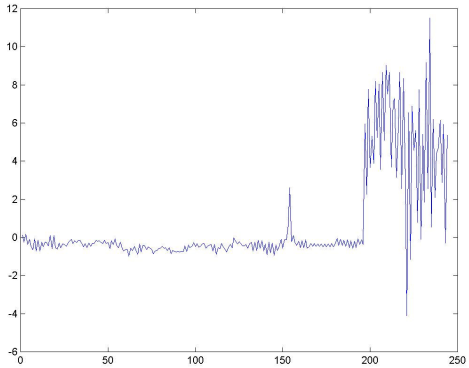

Fig 6. The error of the mapping

The '+' signs in Fig 5 represents the shape of a face showing AU4 while the 'box' symbol represents the shape of the face when showing a neutral expression. It can be seen that the '+'s of the eyebrows are lowered in the image. This shows that it is possible to map a subtle difference in shape without affecting the rest of the shape.

There is a small error here as only the largest three eigenvalues were used in the mapping function, this means that only 85.05% of the total variance was taken into consideration. This can be seen more clearly using Fig 6. The greatest variance is with the points that mark the shape of the eyebrows; these points are around the 200 mark on the x-axis. This provides a measure of the error; it shows the deviation from the original image. Fig 6 is result of subtracting an original neutral image from a synthetic image portraying AU4 of the same individual. The purpose of this image is to show how this model captures the holistic nature of facial expressions. Although only the eyebrows move under this transformation, this model uses the entire face to calculate how they move.

5.2 **Experiment two**

Thirty images of different subjects were used to create a mapping from a neutral expression to a smile. Images were acquired using the same method as experiment one and using the 'AR face database' developed by Martinez [17]. The lighting in these images was constant and the pose and orientation of the images were kept the same, as in experiment one.

The procedure for computing the mapping function was the same as experiment one. As some of the faces were taken from a database that was not consistent with the FACS system, the mapping from a neutral expression to one showing a smile was expected to be less accurate than the mapping generated in the first experiment. That is the intensity of the smile in the images in the training set was not uniform. Fig 7 shows this mapping and its discrepancy.

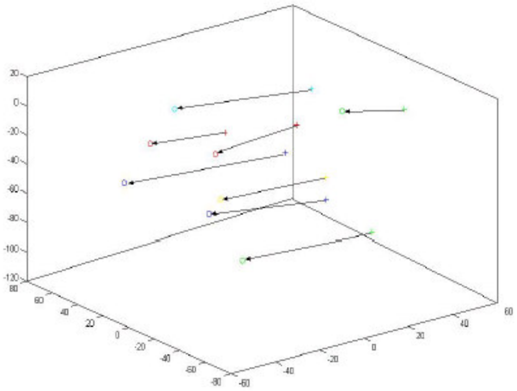

Fig 7. The Mapping taken from a neutral expression to a smile

The '+' sign represents the input vector (the largest three eigenvalues used to portray AU0 or a neutral face for a specific identity) and the 'o' symbol represents the output vector (the largest three eigenvalues used to portray a smile for a specific identity). The non uniform mapping of a face moving from a neutral expression to a smile can be accounted for in two ways, firstly, the images used in the training set were not consistent with the FACS system and secondly, only the first three eigenvalues are represented in the diagram. The first three eigenvalues account for only 65.69% of the total variance found during the training phase. Fig 8 shows how a larger training set increases the amount of components needed to explain the total variance of the training set.

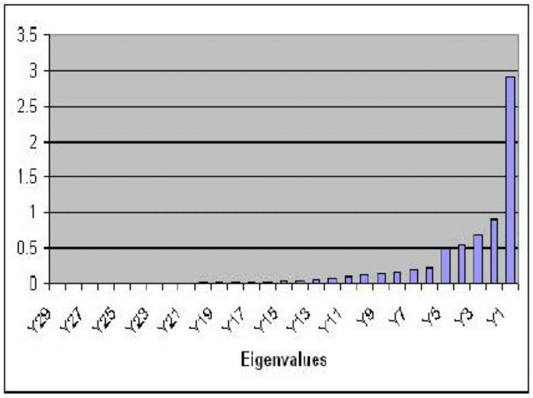

Fig 8. Top 30 eigenvalues

The purpose of experiment two was to demonstrate that a mapping function can be created that can predict accurately how a facial shape will change as the formation of a smile emerges. Fig 9 shows how three never seen before facial shapes would change as predicted by the mapping function as they move from AU0 to a smile. The mapping function was built up using the same FHMN used in experiment one.

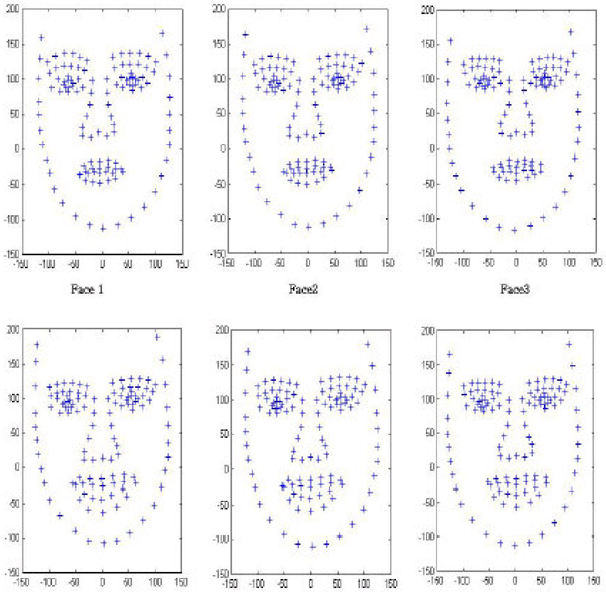

Fig 9. Faces on the bottom row are the results of applying the mapping function to the faces on the top row

The results shown in Fig 9 suggest that any expression can be mapped using the novel technique described in this paper. The images in the top row are representations of the shapes of the never seen before faces and the images in the bottom row are how the images look after been passed through the mapping function which takes an image from a neutral expression to a smile.

6 Concluding remarks and future work

This paper details a method of generating a mapping function as a face moves from a neutral expression to a non-neutral expression. This new method combines the *Facial Action Coding System* (FACS), *Active Shape Model* (ASM) and Artificial Neural Networks (ANN) namely *Feedforward Heteroassociative Memory Networks* (FHMN). The procedure involves the creation of a database of images of faces, the images would portray neutral expressions and non-neutral expressions. The non-neutral expressions selected for a specific mapping would all *score* the same AU's as stipulated by the FACS. All the images would be used in a training phase to build an ASM. The largest principle axes used to portray a neutral expression would be used as input to a FHMN and the largest principle axes used to portray a specific AU would be used as output to a FHMN. In this manner a mapping function is created.

The problem of generating a mapping function as an expression moves from a neutral expression to a non-neutral expression can be simplified through the use of the Active Shape Model (ASM). There were eight modes of variation in the example in experiment one in this paper and the three largest principle components accounted for 85.05% of the total variance. These three principle components were used in a neural network to firstly, see if a mapping exists and secondly, to create a mapping function, if possible. Fig 4 suggests that there is a clear mapping of expression as a face went from a neutral expression to one showing AU4 (brow lowerer), The mapping function was generated and was applied to the inputs. The results show a clear lowering of the eyebrow in each face shape (see Fig 5). This strongly suggest that using the principle components of the shape model as inputs to a trained neural net, any AU can be modelled and therefore a function for expression formation can be created.

In experiment two a larger training set was used and twenty of the images were not consistent with a specific AU as described by the FACS system. The results of this experiment demonstrate two things, firstly that without a proper measure of expression such as the FACS system, this mapping function would be less accurate as suggested by Fig 7, and secondly, that the mapping can be successfully used on never seen before faces (Fig 9). Without a large enough eigenspace to represent a new face it would be impossible to generate the parameters for that facial shape (see equation 9) and hence produce its most probably change in shape.

These functions would allow for classification and generation of expression once a neutral image of a face shape was known. It is predicted that a greater accuracy would be achieved if more examples where used in the training set.

Building the ASM and the ANN to model expression is a computationally complex problem, but because most of the work takes place in the training phase, this new hybrid method consisting of the FACS, the ASM and neural networks can be implemented in real-time. This means that this technique can be used in adaptive applications to create and run virtual avatars and also to identify the users expression. It is important to note that this technique would not classify emotions as a further time dependent system would need to be in place before trying to classify emotions.

It is planned to develop this technique to take into account the grey scale level of the faces as well as the shape, this would be achieved by performing PCA on the shape, the grey scale level and the by performing statistical analysis on how the shape changes with respect to the grey scale levels (this technique is better known as the Active Appearance Model (AAM), Cootes [5])

It is also planned to analyse the trajectory of the expression vector in feature space as an expression changes. This will be done using a training set of people showing different intensities of an AU as described by the FACS system. This will show if the principle components take a specific path through eigenspace as an expression is forming.

References

[1] Anton, and Rorres, “Elementary Linear Algebra”, Seventh Edition, Wiley and Sons, 1994.

[2] Balke and Isard “Active Contours, The Application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion”, Springer, 1998.

[3] Cootes, Taylor, Cooper, and Graham, “Active Shape Models - Their Training and Application”, Computer Vision and Image Understanding, Vol. 61, No. 1, January, pp. 38-59, 1995.

[4] Cootes and Taylor “Statistical Models of Appearance for Medical Image Analysis and Computer Vision”, Imaging Science and Biomedical Engineering, University of Manchester, UK, Proceedings, SPIE Medical Imaging, 2001.

[5] Cootes, Edwards, and Taylor “Active Appearance Model”, 5th European Conference on Computer Vision, Burkhardt and Neumann, eds., Vol. 2, pp. 484-498, Springer, Berlin.1998.

[6] Cootes and Taylor “Statistical Models of Appearance for Computer Vision”, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, U.K. October 26th, 2001.

[7] Cristianini, and Shawe-Taylor “Support Vector Machines”, Cambridge University Press, 2000.

[8] Ekman and Friesen “Facial Action Coding System”, Human Interaction Laboratory, Dept. of Psychiatry, University of California Medical Centre, San Francisco, Consulting Psychologists Press, Inc. 577 College Avenue, Palo Alto, California 94306, 1978.

[9] Faigan “The Artist's guide to Facial Expressions”, Watson-Guphill Publications, 1990.

[10] Forsyth and Ponce “Computer Vision: A Modern Approach”, Prentice Hall, 1st Edition, August 14, 2002.

[11] Ghent, McDonald, and Harper, “A Statistical Model for Expression Generation using the Facial Action Coding System”, NUIM, TR No:NUIM-CS-TR2003-02, 2003

[12] Gladilin, Zachow, Deuflhard, Hege, “Towards a Realistic Simulation of Individual Facial Mimic”, Zuse-Institute-Berlin (ZIB) Takustr, 7, D-14195 Berlin, Germany, November 21-23, 2001.

[13] Gosselin and Schyns “Bubbles: a technique to reveal the use of information in recognition task”, Vision Research, Department of Psychology, University of Glasgow, 56, Hillhead Street, Glasgow G12 8QB, Scotland, UK, December 3rd, 2000.

[14] Gower “Generalised Procrustes Analysis”, Psychometrika, Vol. 40, pp. 33-50, 1975.

[15] Johnson and Wichern, “Multivariate Statistics, A Practical Approach”, Chapman and Hall, 1988.

[16] Lanitis, Taylor and Cootes, “A Generic system for classifying variable objects using flexible template matching”, Proceedings, British Machine Vision Conference, 1993 (J. Illingworth, Ed.), Vol 1, pp 329-338, BMVA Press, 1993.

[17] Martinez, and Benavente, “The AR face database” CVC Tech. Report No.24, 1998.

[18] McDonald, Markham, and McLoughlin, “Selected Problems in Automated Vehicle Guidance”, Signals and Systems Group, Department of Computer Science, NUI Maynooth, 27th April 2001.

[19] Patterson “Artificial Neural Networks, Theory and Application” Prentice Hall, 1995.

[20] Picard, “Affective Computing” MIT Press, 1998.

[21] Principe, Euliano, and Lefebvre, “Neural and Adaptive Systems”, Wiley,2000.

[22] Stephens “Visual Basic Graphics Programming, Second Edition”, Wiley, 1999.

[23] Vernon “Machine Vision, Automated Visual Inspection and Robot Vision”, Prentice Hall, 1991.

[24] Xue, Li, and Teoh, “Facial Feature Extraction and Image Warping using PCA Based Statistical Mode”, Microsoft Research China, Zhi Chun Road, Haidian District Beijing, 100080, P.R. China, 2000