GENERATING A MAPPING FUNCTION FROM ONE EXPRESSION TO ANOTHER USING A STATISTICAL MODEL OF FACIAL TEXTURE

John Ghent
Computer Science Department,
NUI Maynooth, Ireland
email: jghent@cs.may.ie

John McDonald
Computer Science Department
NUI Maynooth, Ireland
email: johnmcd@cs.may.ie

Abstract

We demonstrate a novel method of generating a mapping function which takes an image of a neutral face to an image of the same subject depicting an alternative expression. It is proposed that this mapping function can be used to automatically generate facial expressions from still images of never seen before faces. This technique draws on the work of Ekman's [8] Facial Action Coding System (FACS), which provides an anatomical basis for measuring facial movement. We use the FACS to generate a *Facial Expression Texture Model* (FETM), which is used in conjunction with several *Artificial Neural Networks* (ANN) to develop a mapping function. We describe this method in detail and provide results which demonstrate the effectiveness of this technique.
Keywords*: Facial expression synthesis, Facial Expression Texture Model (FETM), Facial Action Coding System (FACS), function approximation*

1 Introduction

The central goal of this paper is to describe the development of a mapping function which manipulates a neutral image of a subject to accurately display a desired expression. The development of this mapping function involves a comprehensive understanding of expression. Facial expressions have been studied by cognitive psychologists [5, 25], social psychologists [10], neurophysiologists [24], computer scientists [8] and cognitive scientist [6].

The model of facial expression described in this paper is Ekman's [10] Facial Action Coding System (FACS). This method of studying facial expressions and emotions depicted by facial expressions is based on an anatomical analysis of facial actions. A movement of one or more muscles of the face is known as an action unit (AU). All expressions can be described using one, or a combination of the AU's described by Ekman.

We achieve expression synthesis by building a statistical model of the AU in question from a number of subjects showing that expression in a training set. The change in texture of each face in the training phase is analysed and used to derive a mapping function, which takes their neutral face to one depicting the new expression.

To decrease the dimensionality of the mapping the variance in texture of each face in the training set is analysed using *Principal Component Analysis* (PCA). This approach can model a large amount of the variance in the training set by using only a few modes of variation or principal components. This representation of expression is known as the expression space. We use the expression space in conjunction with *Feedforward Heteroassociative Memory Networks* (FHMN) and *Radial Basis Functions* (RBF) to generate a subject independent mapping function and present the results in this paper.

2 Measuring expression

Few studies have measured how the face moves as an expression forms [19, 12, 10, 2, 29]. The central reason for this is the fact that research focused on facial expressions is limited due to the lack of adequate techniques for measuring the face. Knowledge of the muscles of the face allows us to characterise exactly what is happening as an expression is emerging. Since everyone's face is different it is difficult to characterise an expression any other way. For this reason a thorough understanding of the face is required prior to devising a scheme for the characterisation and measurement of facial expression. FACS.

According to Faigin [11], of the twenty-six muscles that move the face, only eleven are responsible for facial expressions. The name of each of these muscles is as follows; (1) Orbicularis oculi, (2) Levator palpebrae, (3) Levator labii superioris, (4) Zygomatic major, (5) Risorius/Platysma, (6) Frontalis, (7) Orbicularis oris, (8) Corrugator, (9) Triangularis, (10) Depressor labii inferioris, and (11) Mentalis. Although this description by Faigin provides a good basis for understanding the anatomy of facial expressions it does not provide an insight as to which muscles work together to create certain expressions.

The *Facial Action Coding System* (FACS) provides a method for studying facial expressions and emotions depicted by facial expressions based on an anatomical analysis of facial actions. A movement of one or more muscles of the face is known as an action unit (AU). Sometimes it is difficult to distinguish if a set of muscles is accountable for a facial movement or if a single muscle is, for this reason the term action unit is used. All expressions can be described using the AU's described by Ekman or a combination of the AU's.

2.1 Facial Expression Texture Model (FETM)

It is necessary that the shape and texture model developed be flexible enough to capture the rules of the FACS but also robust enough to ensure the model can only deform in ways consistent with the training set and in doing so uphold the rules of the FACS. A number of computational techniques exist for building flexible shape models, such as *Hand crafted models* [30, 22, 18] and *articulated model* [1, 17] However, the two most common techniques for representing shapes are $active\ contour\ models$ or $snakes$ and the $Fourier\ series\ shape\ model$. $Active\ contour\ models$ or $snakes$ have been demonstrated to be very effective in generating shape models[3]. These energy minimizing curves are modelled as having stiffness and elasticity and are attracted toward features such as lines and edges. Staib and Duncan [27] use the $Fourier\ series\ shape\ model$ technique effectively to describe medical images and Bozma and Duncan [4] show how this technique can be used to model organs. The central drawback to this technique is the fact that the Fourier transform is only capable of representing band-limited signals. Many contours we deal with are not smooth i.e. they contain corners and hence would require an infinite number of Fourier terms to represent the shape.

To calculate the *Facial Expression Texture Model* (FETM) we warp all images to the mean shape. This is achieved using Delauney triangulation to segment the mean shape into 214 separate triangles using 122 landmark points. We apply the affine transformation to the pixels within each triangle. Suppose $\mathbf{x_1}$, $\mathbf{x_2}$ and $\mathbf{x_3}$ are three corners of a triangle. Any internal pixel can be written as

$$\mathbf{x} = \mathbf{x_1} + \beta(\mathbf{x_2} - \mathbf{x_1}) + \gamma(\mathbf{x_3} - \mathbf{x_1}) = \alpha\mathbf{x_1} + \beta\mathbf{x_2} + \gamma\mathbf{x_3} \tag{1}$$

where $\alpha = 1 - (\beta + \gamma)$ and $\alpha + \beta + \gamma = 1$. For $\mathbf{x}$ to be inside a triangle, $0 < \alpha, \beta, \gamma < 1$. Under the affine transformation, this pixel maps to

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) = \alpha\mathbf{y_1} + \beta\mathbf{y_2} + \gamma\mathbf{y_3} \tag{2}$$

Each face is then converted from colour to greyscale and each image is represented as a vector.

Definition $\mathbf{x}_i^{k_j}$ *Let* $\mathbf{k}$ *be a vector of AU's where* $\mathbf{k} = \{k_0, k_1, k_2 ... k_{m-1}\}$ *and* m *is the number of AU's. Then* $\mathbf{x}_i^{k_j}$ *is a vector representing an image of subject* i *showing AU* k_j.

We use PCA to analyse how the vectors change with respect to each other. Before any significant analysis can be done on the shape of the faces, the mean must be computed. This is done using the equation below:

$$\overline{\mathbf{x}} = \frac{1}{Nm} \sum_{i=1}^{N} \sum_{j=0}^{m-1} \mathbf{x}_i^{k_j} \tag{3}$$

where $\overline{\mathbf{x}}$ is the mean image vector of every subject i portraying every AU k_j and N are the number f subjects in the training set . The difference vector is then calculated using

$$\delta \mathbf{x}_i^{k_j} = \mathbf{x}_i^{k_j} - \overline{\mathbf{x}} \tag{4}$$

where $\delta \mathbf{x}_i^{k_j}$ is the difference between $\mathbf{x}_i^{k_j}$ and the in the mean vector $\overline{\mathbf{x}}$. The covariance matrix is then calculated. In the experiments in this paper the covariance matrix is very large n x n where $n = 65025$ so the eigenvectors and eigenvalues are calculated from a smaller s x s matrix derived from the data, where $s = N$ x m. Let $D = (\delta \mathbf{x}_1^{k_0} \cdots \delta \mathbf{x}_N^{k_m})$. The covariance matrix can be written as

$$S = \frac{1}{s} D D^T \tag{5}$$

Let T be the s x s matrix

$$T = \frac{1}{s} D^T D \tag{6}$$

Let e_i be the s eigenvectors of T with eigenvalues λ_i. The s vectors De_i are all eigenvectors of S with eigenvalues λ_i. All remaining eigenvectors of S have zero eigenvalues. Texture parameters for $\mathbf{x}_i^{k_j}$ can be extracted and reconstructed using a similar technique used with the *Facial Expression Shape Model* (FESM) [13, 16].

3 Function approximation

ANNs have proven to be successful in many practical problems. It has been shown that ANNs can recognise handwritten characters [21], spoken words [20] and more relevantly human faces [9]. In this section we address the problem of facial expression synthesis and discuss ANNs that can be used for this task in conjunction with FETM.

A Feedforward Heteroassociative Memory Network (FHMN) can be used to compute a mapping from x to y. This is a one-layer network that stores patterns and is the simplest type of network we consider.The Neural Network is trained by using the x principal components that represent a neutral face as input and the x principal components that represent a face depicting a specific expression as output. In this manner a mapping function is learned which maps the texture of a neutral face to that of a specific expression.

Radial Basis Function (RBF) networks are a form of ANN that are closely related to what is known as *distance-weighted regression*. The potential of RBF networks has been demonstrated several times [26, 23], In a RBF network each hidden unit produces an activation determined by a radial function (usually a Gaussian) centred at a specific position. In RBF's the learned hypothesis is a function of the form

$$\hat{f}(x) = w_0 + \sum_{u=1}^{k} w_u \mathbf{G}_u(d(x_u, x)) \tag{7}$$

where $\mathbf{G}_u(d(x_u, x))$ is the kernel function. It is common in practice to choose each function $\mathbf{G}_u(d(x_u, x))$ to be a Gaussian function centered at the point x_u. An overview of expression synthesis can now be achieved using the following algorithm.

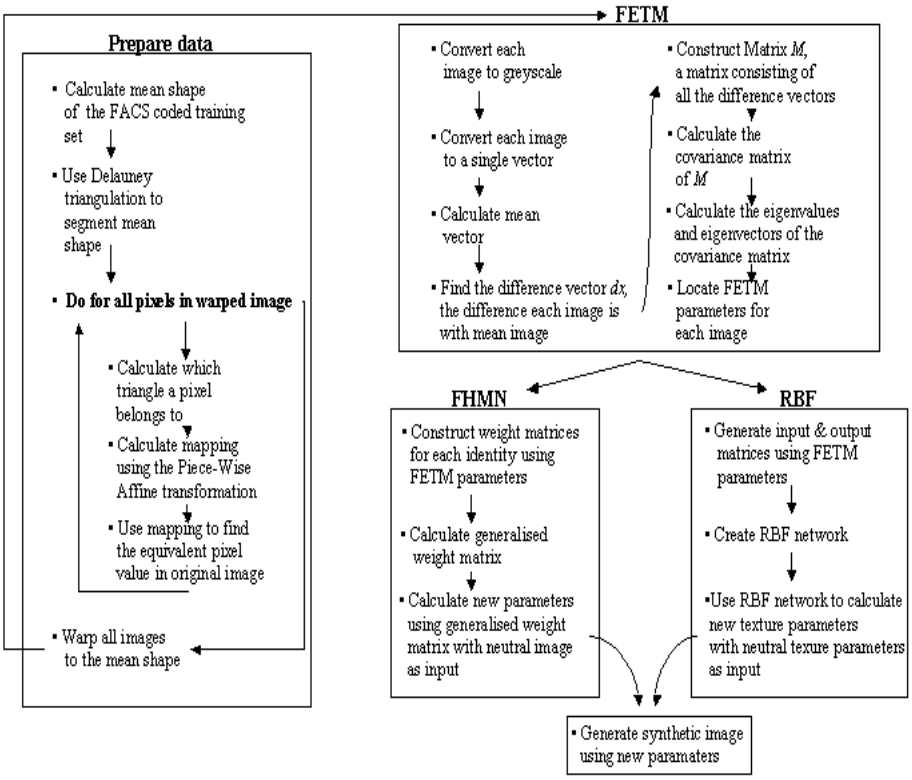

Figure 1: Texture Synthesis

4 Experiments and results

To create a FETM it is necessary to use a database that is consistent with the FACS description of an expression. For this reason we use the Cohn-Kanade AU-Coded Facial Expression Database [7]. The database includes approximately 2000 image sequences from over 200 subjects. All images used from the database are AU coded by certified FACS coders. The images used during the training phase of all experiments described in this paper have been coded as AU 6 + AU 12 + AU 25. A short description of each is provided.

1. **AU 6:** Draws the skin from the temple and cheeks towards the eye. The outer band of muscles around the eye constricts.
2. **AU 12:** Pulls the corners of the lips back and upward, creating a smile shape to the mouth.
3. **AU 25:** Pulls the lips apart and exposes the lips and gums.

Forty people and 80 images from the Cohn-Kanade AU-coded facial expression database were used. Each image was acquired using a Panasonic WV3230 camera connected to a Panasonic S-VHS AG-7500 video recorder. The camera was located directly in front of the subject, and each image was digitized into 640 by 480 pixel arrays of greyscale values.

The mean shape was segmented using Delauney triangulation and each image was warped to the mean shape using a piece-wise affine transformation. The mean image was then calculated (Fig 2). Each image was then represented as a single vector, subtracted from the mean image and the FETM was generated. The top 30 principal components of the FETM describe 95.60% of the total variance found in the training set. Fig 3 illustrate the effect of varying the top four principle components.

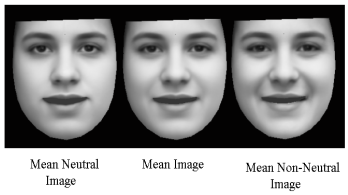

Figure 2: The mean images

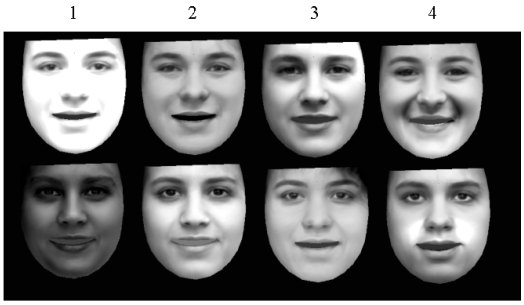

Figure 3: Top four principal components

A FHMN was used to generate a mapping from a neutral expression to one depicting AU 6, AU 12 and AU 25. Of the 40 subjects used to create the FETM, 37 subjects were used during the training of network. This network generalized the mapping too much the identity was lost when implemented with the FETM. Fig 4 emphasizes this generalization. It should be noted that the change in shape in Fig 4 is calculated using the *Facial Expression Shape Model* (FESM) [13, 15, 16].

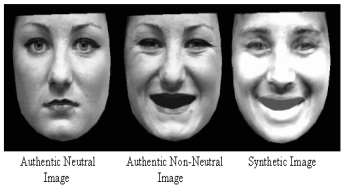

Figure 4: Expression Synthesis using a FHMN

To improve the mapping further we used a more sophisticated *Radial Basis Function Network* (RBFN) with the FETM. The top 30 principal components of the FETM were used to train the RBF. The training data consisted of 37 subject and 74 images. Three subjects were excluded from the training of each network to test each network with unseen data. The table below shows the correlation coefficients between the estimated and real principal components for the FETM in conjunction with a RBF network.

$Table_1$	$Experiment_1$	
$Subject$	$FETM$	RBF
1	0.9999	
2	1	
3	0.6017	
4	0.6771	
5	0.6208	
$Average$	0.7799	

Subjects one and two were used with 35 other subjects to train the network while subjects three, four and five are unseen test data. The test data for the FETM has a correlation coefficient of $t_{avg} = 0.6645$.

Using a similar technique Yangzhou and Xueyin [28] showed how a $uniform\ function$ achieves results of $a_{avg} = 0.51$. This technique improves on this by computing a uniform function that achieves considerably better results. Fig 5 shows the error of the mapping within the FETM, the histogram on the left is the error of the mapping for all images in the training set and the histogram on the right shows the error for all the unseen images. Fig 6 illustrates the photo-realistic synthetic facial expressions of five different subjects. The first two rows consists of images of subjects that were used during the training of the RBF network while the next three individuals (rows 3, 4 and 5) were not used during the training of the network. Column one consists of shape free original images of individuals depicting neutral expressions, column two consists of shape free original images of individuals depicting AU 6, AU 12 and AU 25 as described by the FACS. Column three consists of synthetic images of individuals portraying AU 6, AU 12 and AU 25 as calculated by the RBF network with neutral image parameters as input and the FETM. Columns 4, 5 and 6 are the same as the first three columns respectively except with shape taken into consideration. The shape in column 6 is calculated using a FHMN in conjunction with the *Facial Expression Shape Model* (FESM) [13, 16].

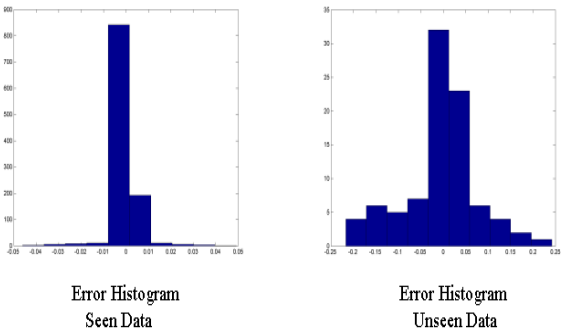

Figure 5: Error of the mapping

Figure 6: Original neutral, original non-neutral and synthesized images.

5 Conclusion and future work

This paper showed how a uniform mapping function was created which toke a neutral image of a face to one depicting a desired facial expression. This was achieved by the development of FETM and using this model several networks were trained to develop an accurate universal mapping function.

The FETM is based on the FACS, an anatomical analysis of facial actions. The FACS provides us with a universal method of analyzing facial expression and allowed for the generation of a texture model that is independent of subject (age, sex, skin colour etc.). The top 30 principal components of the FETM could describe 95.60% of the total variance found in the training set.

A FHMN was used to develop mapping functions which toke an image of a neutral face to one depicting a smile (AU 6, AU 12, AU 25). This network over generalized the mapping and hence much of the identity of a subject was lost during the calculations. To improve the results a more sophisticated RBF network was used with the FETM. This networks greatly improved the results and a correlation coefficient between synthesized and authentic images of $t_{avg} = 0.6645$ was achieved. The results can be seen more clearly in Fig 6. The first two rows of this diagram show expression synthesis on data that was used during the training phase, this diagram shows how this technique successfully differentiates between skin colour. The images in the last three rows are images that were not present during the training phase. These images illustrate how this technique can generate a synthetic expression of a subject regardless of sex.

It is planned to use the FETM for expression classification. This could be done using similar neural networks to the ones detailed in this paper.

References

[1] Beinglass, A. and Wolfson, H. J. "Articulated object recognition", Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 461-466, 1991

[2] Birdwhistell, R.I "Kinesics and Context" Philadelphia: university of Pennsylvania Press, 1970.

[3] Balke, A and Isard, M, "Active Contours, The Application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion",(Springer, 1998).

[4] Bozma, H. I. and Duncan, J. S. "Model-based recognition of multiple deformable objects using a game theoretic framework", Information Processing in Medical Imaging-Proceedings of the 12th International Conference, pp. 358-372, Springer-Verlag, Berlin/New York, 1991

[5] Bruce, V. Young, A. "Understanding face recognition". British Journal of Psychology, 77: 305-328. 1986.

[6] Brunelli, R. Poggio, T. "Face Recognition Features versus Templates" IEEE Transactions on PAMI, 15(10): 1042-1052, 1993.

[7] Cohn, J. Kanade "Cohn-Kanade AU-Coded Facial Expression Database", Pittsburgh University, 1999.

[8] Cootes, T. F. and Taylor, C. J. "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.

[9] Cottrell, G.W. Metcalfe, J. "Face, emotion and gender recognition using holons" Proceedings of the 1990 conference on advances in neural processing systems 3, 564-571, 1990.

[10] Ekman, P. and Friesen, W. V. "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.

[11] Faigan, G. "The Artist's guide to Facial Expressions", Watson-Guphill Publications, 1990.

[12] Fulcher, J.S. “Voluntary facial expressions in blind and seeing children. ” Archives of Psychology, ,38(272), 1942.

[13] Ghent, J. McDonald, J. “Generating a Mapping Function from one Expression to another using a Statistical Model of Facial Shape”, Proceedings of the Irish machine vision and image processing conference, 2003

[14] Ghent, J. McDonald, J and Harper, J. “A Statistical Model for Expression Generation using the Facial Action Coding System”, NUIM, NUIM-CS-TR2003-02, technical report, Jan 2003

[15] Ghent, J. McDonald, J. “An Overview of a Computational Model of Facial Expression”, NUIM postgraduate symposium, March 2004.

[16] Ghent, J. McDonald, J.“A Computational Model of Facial Expression”, NUIM-CS-TR-2004-01, technical report, Jan 2004.

[17] Grimson, W. E. L., “Object Recognition by Computer: the Role of Geometric Constraints”, MIT Press, Cambridge, MA, 1990

[18] Hill, A. and Taylor, C. J. “Model based image interpretation using genetic algorithms”, Image Vision Comput. 10, pp. 295-300, 1992

[19] Landis, C. “Studies of emotional reactions: II. General behavior and facial expressions” Journal of Comparitive Psychology, 4:447-509, 1924

[20] Lang, B. “The effects of processing requirments on neurophysiological responces to spoken sentences” PubMed 12191461 39(2): 302-318, 1990

[21] LeCun, Y. Boser, B. Denker, J.S. Henderson D. Howard, R.E. Hubbard, W. Jackel, L.D. “Backpropagation applied to handwritten zip code recognition” Neural Computation, 1(4): 541-551, 1989.

[22] Lispon, P. Yuille, A. L. O’Keeffe, D. Cavanaugh, J. Taaffe, J. and Rosenthal, D. “Deformable templates for feature extraction from medical images”, Proceedings of the first European Conference on Computer Vision (O. Faugers, Ed.), Lecture notes in Computer Science, pp. 413-417, Springer-Verlag, Berlin/New York, 1990

[23] Moody, J. Darken, C. “Fast learning in Networks of locall-tuned processing units” Neural Computation, 1:281-294, 1989.

[24] Perret, M. Hietanen, J.K. Oram, P. Benson, P. “The effects of lighting conditions on response of cells selective to face views in the macaque temporal cortex” Exp. Brain Res. 89: 157-71, 1992.

[25] Rhodes, G. Brake, S. and Atkinson, A. “Whats lost in inverted faces?” Cognition, 47: 25-57, 1993.

[26] Powell, J.D. “Radial basis functions for multivariate interpolation: a review” Clarendon Press, Oxford, UK, 1986.

[27] Staib, L. H. and Duncan, J. S. “Parametrically deformable contour models”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, pp 427- 430, 1989

[28] Yangzhou, D. Xueyin, L. “Emotional facial expression model building”, Pattern recognition letters 24, pp 2923-2934, 2003

[29] Young. G and Decarie, T.G. “An ethology-based catalogue of facial/vocal behaviours in infancy” Archives of Psyhology. 37, No. 264, 1941.

[30] Yuille, A. L. Cohen, D. S. and Hallinan, P. “Feature extraction from faces using deformable templates”, Int. J. Comput. Vision 8, 99-112, 1992