Available online at www.sciencedirect.com
SCIENCE DIRECT
Image and Vision Computing 23 (2005) 1041–1050
image AND vision COMPUTING
www.elsevier.com/locate/imavis

Photo-realistic facial expression synthesis

John Ghent*, John McDonald

Department of Computer Science, National University of Ireland Maynooth, Ireland

Received 15 September 2004; received in revised form 27 May 2005; accepted 15 June 2005

Abstract

This paper details a procedure for generating a function which maps an image of a neutral face to one depicting a desired expression independent of age, sex, or skin colour. Facial expression synthesis is a growing and relatively new domain within computer vision. One of the fundamental problems when trying to produce accurate expression synthesis in previous approaches is the lack of a consistent method for measuring expression. This inhibits the generation of a universal mapping function. This paper advances this domain by the introduction of the Facial Expression Shape Model (FESM) and the Facial Expression Texture Model (FETM). These are statistical models of facial expression based on anatomical analysis of expression called the Facial Action Coding System (FACS). The FESM and the FETM allow for the generation of a universal mapping function. These models provide a robust means for upholding the rules of the FACS and are flexible enough to describe subjects that are not present during the training phase. We use these models in conjunction with several Artificial Neural Networks (ANN) to generate photo-realistic images of facial expressions.

Keywords: Image synthesis; Facial expression shape model (FESM); Facial expression texture model (FETM); Radial basis function network (RBFN)

1. Introduction

Facial expressions play a major role in how people communicate. They serve as a window to one's own emotional state, they make behaviour more understandable to others and they supplement verbal communication. A computer that could interact with humans through facial expression would advance human–computer interfaces and provide a basis for communication that could be compared to human–human interaction.

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. There has been a lot of work in this area over the past 5 years [1–5]. However, real-time photo-realistic expression synthesis for unseen images has not yet been achieved [6]. The approach presented in this paper combines the FACS, statistical shape and texture models, and machine learning techniques to provide a novel solution to this problem.

In the past Radial Basis Function Networks (RBFN) have been applied to facial expression synthesis [7,8]. However, in these approaches no redundancy reduction techniques were applied prior to calculating the mapping functions. This kept the dimensionality of the mapping functions high and meant that irrelevant information was used in calculating the mapping functions. In King's [7] approach mapping functions were used to modify the locations of Facial Characteristic Points (FCP) which in turn were used to warp an image to depict an alternative expression. A key weakness with this approach is that in order to adequately model the appearance change due to expression, one must take account of the variation of both shape and texture. For example, to synthesis a smile the texture of the image must be modified to produce wrinkles. The technique described in this paper overcomes this problem by manipulating both shape and texture of the input image.

More recently, Abboud [9] applied PCA to the shape and texture of unseen images to lower the dimensionality of the problem in order to produce synthetic facial expressions. However, the approach used linear regression to perform facial expression synthesis. Our technique improves on this approach by using a RBFN to describe the non-linear nature of facial expressions. The results of

* Corresponding author. Tel.: +353 1 7084533; fax: +353 1 7083834.
E-mail address: jghent@cs.may.ie (J. Ghent).

0262-8856/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.imavis.2005.06.011

the technique described in this paper considerably outperform the results found in [9].

1.1. Overview

This paper details a technique that allows for real-time photo-realistic images of a person depicting a desired expression to be synthesised once a neutral image of the subject is present [10,11]. The development of this mapping function involves a comprehensive understanding of expression. In the past facial expressions have been studied by cognitive psychologists [12,13], social psychologists [14], neurophysiologists [15], computer scientists [16] and cognitive scientists [17].

As everyone's facial features are unique, it is hypothesised that the only feasible way to measure an expression is not by examining features of expressions, but by the movement of the muscles of the face. For this reason, the anatomy of the face is a very important aspect in understanding expression. The model of facial expression described in this paper is Ekman's Facial Action Coding System (FACS) [14]. 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 AUs described by Ekman. In previous approaches, the training set has caused inaccuracies in the mapping functions [7]. This paper solves this problem by only using images that are FACS coded to a specific expression to be included in the training set.

We achieve expression synthesis by building statistical models of the AUs in question from a number of subjects showing that expression in a training set. These models must provide adequate flexibility to cover the differences in human facial expression, however, it is also necessary that the model only deforms in a manner consistent with the system used to measure facial expressions. The change in shape and texture of each face in the training phase is analysed and used to derive a mapping function, which maps an image of their neutral face to one depicting a new expression.

To decrease the dimensionality of the mapping, the variance in the shape and 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), Linear Networks (LN) and Radial Basis Function Networks (RBFN) to generate subject independent mapping functions.

Photo-realistic facial expression synthesis could be used in numerous applications. A short-list is detailed below:

- To generate arbitrary animated agents.
- To automate interactive web hosts for low bandwidth video conferences.
- To add personality and expressions to arbitrary images of faces.
- For interactive games.
- For biometric systems that are invariant to expressions.
- Viseme (visual equivalent of phonemes) synthesis.

The structure of this paper is as follows: Section 2 details facial expressions. Section 3 details the construction of the FESM and the FETM model. Section 4 describes the learning phase of this technique, Section 5 details experimental results with our proposed techniques and section six provides some concluding remarks and future work.

2. Facial expressions

Relatively few studies have measured how the face moves as an expression forms [14,18–20]. The central reason for this is due to the lack of adequate techniques for measuring the face. More recent approaches to facial measurement have varied in methodology, from measurements of specific changes to a particular part of a face [19], to verbal descriptions of facial gestalts [20]. Knowledge of

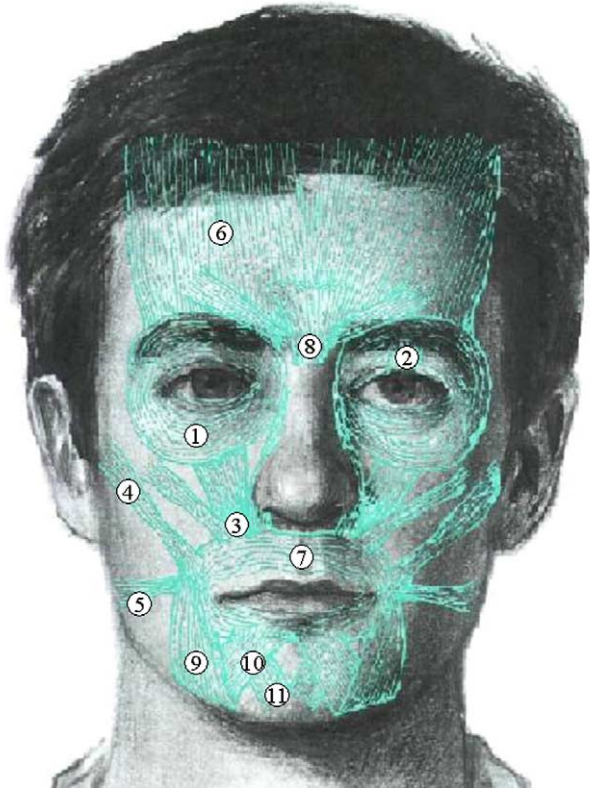

Fig. 1. The 11 most influential muscles in expression formation (used by permission [21]).

the muscles of the face allows us to characterise exactly what is happening as an expression is emerging. Since the appearance of everyone's face is different it is difficult to characterise an expression any other way. For this reason, a thorough understanding of facial anatomy is required prior to devising a scheme for the characterisation and measurement of facial expression. In Section 2.1, we describe the elements of the facial anatomy responsible for the formation of facial expression. Subsequent to this we detail the FACS.

2.1. Anatomical analysis

According to Faigin [21], of the 26 muscles that move the face, only 11 are responsible for facial expressions. These muscles are shown in Fig. 1 and consist of: (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..

2.2. The Facial Action Coding System (FACS)

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). It can be difficult to distinguish if a single muscle or a set of muscles is accountable for a facial movement, for this reason the term action unit is used. All expressions can be described using the AUs described by Ekman [14].

2.3. Scoring action units

For an AU to be considered present, there has to be a significant appearance change. Appearance changes are characterised by the following criteria: (1) the parts of the face that have moved and the direction of their movement, (2) the wrinkles that have appeared or have become more pronounced, and (3) the alterations in the shape of the face. The intensity of an AU can also vary from *trace* to *maximum*. There are five intensities in total ranging from *A*, minimum to *E*, maximum. These are shown in the following list.

A. *Trace*. Activity in the face is barely noticeable.
B. *Slight*. One or more changes in appearance are visible.
C. *Marked/pronounced*. Distinguished from B by a set of criteria that is specific to each AU. This criteria establishes how much more evidence is required to score C.

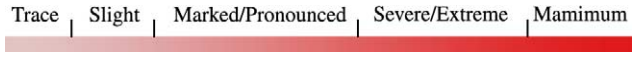

Fig. 2. The scale of intensity scores.

D. *Severe/extreme*. Similar to the distinction between B and C.
E. *Maximum*. Maximum appearance change required to score E.

Fig. 2 provides a visual representation of the relationship between the scale of evidence for classification and the intensity scores. Fig. 2 can be thought of as the spectrum of intensity for each AU. Note that this scale is divided non-uniformly, e.g. *marked/pronounced* takes up a larger range of the change than trace.

3. Shape and texture models

A number of computational techniques exist for building flexible shape models. *Hand crafted models* can be developed using circles, lines, and arcs that can move around relative to one another. Yuille et al. [22] demonstrated this technique in modelling parts of the face such as the eyes and mouth. Lipson et al. [23] and Hill et al. [24] illustrated the usefulness of this technique by building an elliptical model of vertebrae and by building a handcrafted model of the heart, respectively. Another useful technique is the *articulated model* which is built from rigid components connected by sliding or rotating joints. Beinglass and Wolfson [25], and Grimson [26] demonstrate the effectiveness of this technique by locating an object within an image.

The two most common techniques for representing shapes are *active contour models* [27] or *snakes* and the *Fourier series shape model* [28]. Active contour models or snakes have been demonstrated to be very effective in this domain. These energy minimising curves are modelled as having stiffness and elasticity and are attracted toward features such as lines and edges. Equilibrium equations allow the curve to move toward image features whilst ensuring the curve's original shape and smoothness are maintained. In this way, the spline moves towards the dominant edge contours and hence the most probable match for its shape in the image.

Hinton et al. [29] illustrates this technique by allowing a set of control points govern the movement of a spline. Each control point has a desired 'home' location which acts as the shape's restoring force, the shape deforms by movement of the control points. The main problems with this technique are, the fact, that the shape is infinitely deformable, and contains no information on whether a shape belongs to a class of shape or not. This creates problems when trying to create a model of facial expressions as a model of this nature has to be based on strict rules.

Staib and Duncan [30] use the Fourier series shape model technique effectively to describe medical images and Bozma and Duncan [31] show how this technique can be used to model organs. The central drawback to this technique is 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. For these reasons, a statistical model based on point distribution is used that only allows deformations observed from the training set and accurately describes the training set. The model adopted is based on Cootes's work on statistical models of appearance [16].

3.1. Facial expression shape model

In order to develop the Facial Expression Shape Model (FESM), we first have to label every image with a set of landmark points. These are located around key areas such as the eyes, nose, mouth and eyebrows (see Fig. 3).

3.1.1. Shape alignment

To analyse the variance of the points that describe the shape of the face it is necessary that the faces in the training set are as closely aligned as possible. One way to achieve this is to use a technique known as Generalised Procrustes Alignment (GPA) [32]. This technique aligns two shapes with respect to position, rotation and scale by minimising the weighted sum of the squared distances between the corresponding landmark points. The alignment depends on the weights given to each of the points, which in turn depends on which AU is being mapped.

To align two shapes **P** and **Q**, we choose a rotation, scale and translation that minimises the sum of the squared distances between them. Let **P** be defined as

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_{n-1}] = \begin{bmatrix} x_1, x_2, x_3, \ldots, x_{n-1} \\ y_1, y_2, y_3, \ldots, y_{n-1} \end{bmatrix} \tag{1}$$

and

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \ldots, \mathbf{q}_{n-1}] = \begin{bmatrix} x_1, x_2, x_3, \ldots, x_{n-1} \\ y_1, y_2, y_3, \ldots, y_{n-1} \end{bmatrix} \tag{2}$$

Procrustes alignment computes the transformation by minimising the error function:

$$E = (\mathbf{P} - \mathbf{MQ} - \mathbf{t})^{\mathrm{T}} \mathbf{W} (\mathbf{P} - \mathbf{MQ} - \mathbf{t}) \tag{3}$$

Here, **M** represents the rotation and the scale of **Q**, i.e.

$$\mathbf{M} = \mathbf{M}(s, \theta) = \begin{pmatrix} (s \cos\theta) - (s \sin\theta) \\ (s \sin\theta) + (s \cos\theta) \end{pmatrix} \tag{4}$$

and **t** is the translation. Hence, Procrustes alignment leads to four linear equations which can be solved using traditional matrix methods. To align a set of N shapes, which is the case in this situation, the following two step algorithm is used:

- Calculate the mean shape
- Align each shape to the mean shape by minimising (3).

This process of alignment is demonstrated in Fig. 4 by a simple example. The Fig. 4 shows the mean shape as a solid line and the variance of other landmark points from the mean.

3.1.2. Principal component analysis

Principal Component Analysis (PCA, also known as the Karhunen–Loéve transform) is a technique used to lower the dimensionality of a feature space [32]. This method

Fig. 3. The arrangement of landmark points around the faces.

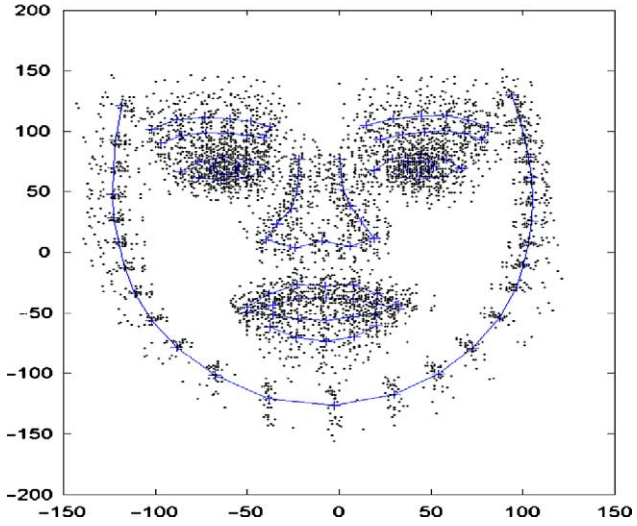

Fig. 4. Aligned landmark points.

takes a set of data points and constructs a lower dimensional linear subspace that best describes the variation of these data points from their mean. We use PCA here to analyse how the landmark points move with respect to each other.

Before any significant analysis can be done on the shape of the faces, the mean shape must be computed. We then take the set of landmark points and constructs a lower dimensional linear subspace that bests describes the variation of these data points from their mean. The covariance matrix is calculated using

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(\delta\mathbf{P}_i)(\delta\mathbf{P}_i^{\mathrm{T}}) \tag{5}$$

where $\delta\mathbf{P}_i$ is the difference between $\mathbf{P}_i$ and the mean shape, and N is the number of shapes is the training set. 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. Images can be reconstructed using

$$\mathbf{P} = \bar{\mathbf{x}} + \mathbf{S}\mathbf{b} \tag{6}$$

where $\mathbf{S}$ is the set of eigenvectors, $\mathbf{b}$ is a vector of weights and $\bar{\mathbf{x}}$ is the mean image. Since images can be approximated using $\mathbf{b}$ (in conjunction with $\mathbf{S}$ and $\bar{\mathbf{x}}$), representing expression in this manner provides for simpler manipulation of expression without any significant loss of information.

3.2. Facial Expression Texture Model

To calculate the Facial Expression Texture Model (FETM) we warp all images to the mean shape. This is achieved using Delaunay triangulation to segment the mean shape into 214 separate triangles using 122 landmark points. We apply an affine transformation to the pixels within each triangle. This is achieved by computing the barycentric coordinates of each point relative to its surrounding triangle in the input image. The output points are identified as the points with equivalent barycentric coordinates in the corresponding triangles in the output image.

It is often the case that there does not exist enough information to give values to every pixel in the output image. This is overcome using bilinear interpolation. It estimates the value of an unknown pixel by using the pixels around it. We use PCA again to analyse how the warped images change with respect to each other. In the experiments in this paper the $n \times n$ covariance matrix is very large, where $n = 65025$. For this reason, the eigenvectors and eigenvalues are calculated from a smaller $N \times N$ matrix derived from the data. Texture parameters can be extracted and reconstructed using a similar technique used with the Facial Expression Shape Model (FESM). For a more detailed description of the construction of the FETM see [11].

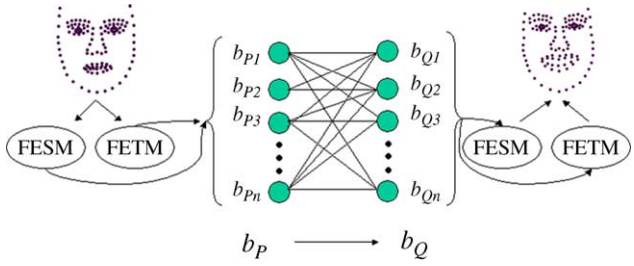

Fig. 5. Structure of mapping function.

4. Learning phase

In this section, we address the problem of facial expression synthesis and discuss ANNs that can be used for this task in conjunction with the FESM and the FETM.

All neural networks are trained by using the parameters that describe neutral faces as input and the parameters that depict a specific expression as target. The structure of these mapping functions is illustrated in Fig. 5.

4.1. Calculating mapping functions

A Feedforward Heteroassociative Memory Network (FHMN) is used to compute a mapping from a neutral expression to one depicting an alternative expression [33]. This is a one-layer network that stores patterns and is the simplest type of network we consider. The synaptic weight value of a FHMN is given by

$$\mathbf{W}_i = \mathbf{b}_p(\mathbf{b}_q)^{\mathrm{T}} \tag{7}$$

where $\mathbf{b}_p$ are the parameters of shape or image p and b_q are the parameters of shape or image q. Eq. (7) is person specific. To generate a general mapping function the following equation is used

$$\mathbf{W} = \alpha\sum_{i=1}^{N}\mathbf{W}_i \tag{8}$$

where α is for normalisation and N is the number of samples in the training data. To produce a synthetic expression, the associated parameters $\mathbf{b}_{\mathrm{p}}$ are used as input to the network, thus

$$\mathbf{b}_{\mathrm{p}} = \mathbf{W}\cdot\mathbf{b}_{\mathrm{p}} \tag{9}$$

$\mathbf{b}_q$ can then be used in conjunction with the FESM and the FETM to generate the shape and texture.

A Linear Network (LN) is a perceptron with a linear output instead of a hard-limiting output. This means that their outputs can take on any value, which is needed for function approximation, whereas the perceptron output is limited to either 0 or 1. Like the FHMN this type of network can only solve linearly separable problems. This network is trained to minimise the error between the input and output training data. This is achieved using the Least Mean Squares (Widrow–Hoff) algorithm [33]. Each input is applied to

the network and the network output is compared to the target. The error is calculated as the difference between the target output and the network output. The error is calculated using:

$$E = \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{a}_i)^2 \tag{10}$$

This is the error of the mapping from a neutral expression to an alternative expression, where $\mathbf{a}_i$ is the output of the network given the input $\mathbf{b}_i$ and $\mathbf{t}_i$ is the target output. The LMS algorithm adjusts the weights and biases of the linear network so as to minimise this mean square error.

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 [34,35], most notably for facial expression recognition [36].

In a RBF network, each hidden unit produces an activation determined by a radial function (usually a Gaussian) centred at a specific position. Neurons are added to the network until the sum-squared error falls beneath an error goal or a maximum number of neurons have been reached. In RBF's, the learned hypothesis is a function of the form

$$\hat{f}(x) = w_0 + \sum_{u=1}^{k} w_u \mathbf{G}_u(d(x_u, x)) \tag{11}$$

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 cantered at the point x_u. The RBF transfer function used in this paper is $t = \mathrm{e}^{-n^2}$. To retrieve parameters of an image $\mathbf{b}_p$ given $\mathbf{b}_q$ we use

$$\mathbf{b}_p = \mathbf{W}_{\mathbf{HLL}}(\mathrm{e}^{-(\|\mathbf{W}_{\mathbf{RBF}} - \mathbf{b}_q\| \mathbf{b}_{\mathbf{RBF}})^2}) + \mathbf{b}_{\mathbf{HLL}} \tag{12}$$

where $\mathbf{W}_{\mathrm{RBF}}$ is the weight matrix of the radial basis layer, $\mathbf{b}_{\mathrm{RBF}}$ is the bias vector of the radial basis layer, $\mathbf{W}_{\mathrm{HLL}}$ is the weight matrix of the hidden linear layer and $\mathbf{b}_{\mathrm{HLL}}$ is the bias vector of the hidden linear layer.

5. Experiments

To create a FESM and 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 [37]. 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 AU6+AU12+AU25, AU1+AU2+AU5+AU26 and AU15+AU17. A short description of AU6+AU12+AU25 is provided, for descriptions of other AUs see [14].

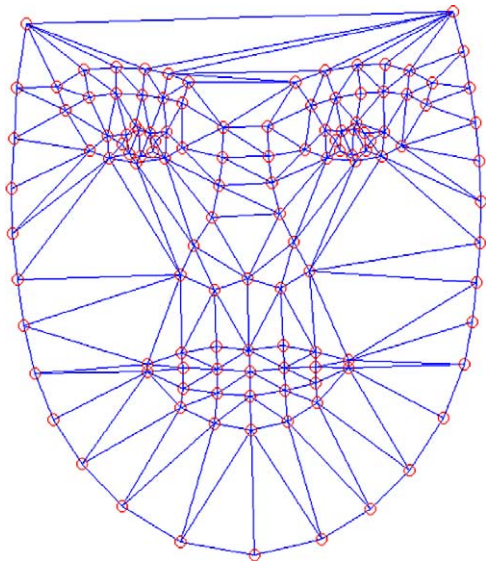

Fig. 6. Mean shape segmented using Delaunay triangulation.

(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.

Seventy-seven people and 154 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 originally digitised into 640×480 pixel arrays of greyscale values.

Each face was manually labelled using 122 landmark points and aligned with each other using Procrustes alignment. PCA was performed on the data and the top few principal components were used in the FESM. The mean shape was segmented using Delaunay triangulation and each image was warped to the mean shape using

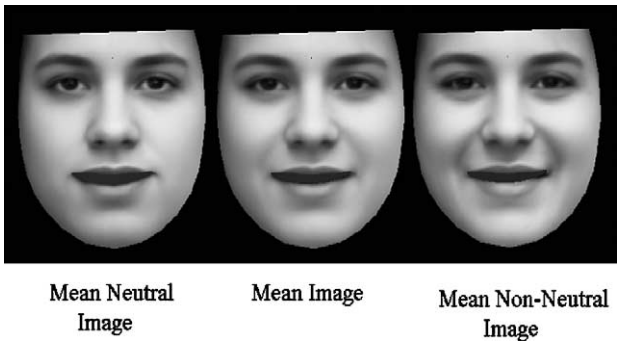

Fig. 7. The mean images.

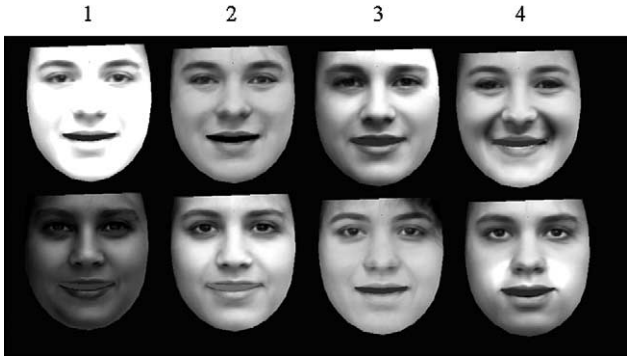

Fig. 8. Top four principal components.

Fig. 9. Expression synthesis using a FHMN.

a piece-wise affine transformation. The segmented mean shape can be seen in Fig. 6.

The mean image was then calculated (Fig. 7). 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 for AU6+AU12+AU25 describe 95.59% of the total variance found in the training set. Fig. 8 illustrates the effect of varying the top four principal components.

A FHMN was used to generate a mapping from a neutral expression to one depicting an alternative expression. This network was implemented on the shape and texture separately to create two mappings. This network, using the FESM produced encouraging results but failed to return convincing results when using the FETM. Fig. 9 illustrates the effect of passing the shape and texture of an image through these mapping functions. The shape and texture parameters were calculated using the neutral image on the left of Fig. 9. These parameters were passed through the neural networks and the image on the right is the output. The target image is at the center of the figure. It can be shown that the identity of the person is lost as the FHMN in conjunction with the FETM performs poorly and while the shape model performs well it is thought that the output can be enhanced by using a more advanced network.

To improve the mapping further, we used a Linear Network (LN) with the FESM and used a more sophisticated Radial Basis Function Network (RBFN) with the FETM. Fig. 10 illustrates the photo-realistic synthetic facial expressions of five different subjects. The first two rows consist of images of subjects that were used during the training of the networks while the next three individuals (rows 3–5) were not used during the training of the networks. The images in the first-three columns are all warped to the mean shape and are therefore considered to be

Fig. 10. Original neutral, original non-neutral and synthesised images.

Table 1
Correlation coefficients between the estimated data and real data of the FESM using a FHMN and a LN

AU	N_I	N_T	Parm	Perc	Seen/ Unseen	N_S	FHMN			LN		
							Avg	Min	Max	Avg	Min	Max
6,12,25	80	40	15	94.04	Seen	35	0.6929	0.3356	0.9433	0.9586	0.7539	0.9976
					Unseen	5	0.5220	0.3353	0.8244	0.8746	0.5896	0.9965
1,2,5,26	40	20	20	98.61	Seen	15	0.7456	0.3842	0.9200	0.9755	0.8990	0.9991
					Unseen	5	0.6019	0.4678	0.8941	0.7773	0.5223	0.9714
15,17	34	17	20	99.35	Seen	15	0.7465	0.3333	0.9184	0.9692	0.7761	0.9993
					Unseen	2	0.6811	0.6286	0.7336	0.6995	0.5536	0.8454
Total	154	77	N/A	N/A	Seen	65	0.7281	0.3333	0.9433	0.9677	0.7539	0.9993
					Unseen	12	0.6017	0.3353	0.8941	0.7838	0.5223	0.9965

Table 2
Correlation coefficients between the estimated data and real data of the FETM using a RBFN

AU	N	Parm	Perc	Seen/unseen	N	Avg	Min	Max
6,12,25	40	30	95.59	Seen	35	99.66	93.55	1
				Unseen	5	77.99	62.80	99.99
1,2,5,26	20	10	89.18	Seen	15	99.07	87.49	99.95
				Unseen	5	77.58	31.66	87.49
15,17	17	10	92.03	Seen	14	97.73	73.67	99.99
				Unseen	3	63.51	50.11	73.67
Total	77	N/A	N/A	Seen	64	98.82	73.67	1
				Unseen	13	73.027	31.66	99.99

shape free. Column one consists of shape free original images of individuals depicting neutral expressions. Column two consists of shape free original images of individuals depicting AU6 + AU12 + AU25 as described by the FACS. Column three consists of synthetic images of individuals portraying AU6 + AU12 + AU25 as calculated by the RBF network with neutral image parameters as input. Columns 4–6 are the same as the first three columns, respectively, except with shape taken into consideration. The shapes in column 6 are calculated using a Linear Network in conjunction with the FESM.

In order to evaluate the performance of this technique we find the correlation coefficient between the estimated data and the real data. Table 1 shows the correlation coefficients between the estimated and real principal components for the FESM using a linear network and a FHMN. N_I is the number of images in the experiments, N_T is the number of identities involved, Parm is the number of principal components used during the training of the neural networks and Perc is the percentage of variance that Parm can describe. N_S is the number of individuals used for training and testing the mapping functions while Avg, Max and Min are the average, maximum and minimum correlation coefficients between the estimated shape parameters and the real shape parameters.

Table 2 shows the correlation coefficients between the estimated and real principal components for the FETM using a RBFN. N_I, N_T, Parm, Perc, Parm and N_S are the same as in Table 1, while Avg, Max and Min are the average, maximum and minimum correlation coefficients between the estimated texture parameters and the real texture parameters. Fig. 11 is of a person who was present during learning phase while Fig. 12 are of people who were not present during learning phase. In Figs. 11 and 12, the shape is calculated using the FESM and a LN while the texture is calculated using the FETM and a RBFN.

It is shown that the average correlation coefficient is $a_{avg} = 0.757$ using the FESM with the LN and the FETM

Fig. 11. Original and synthesised image of seen data.

Fig. 12. Original and synthesised images of unseen data.

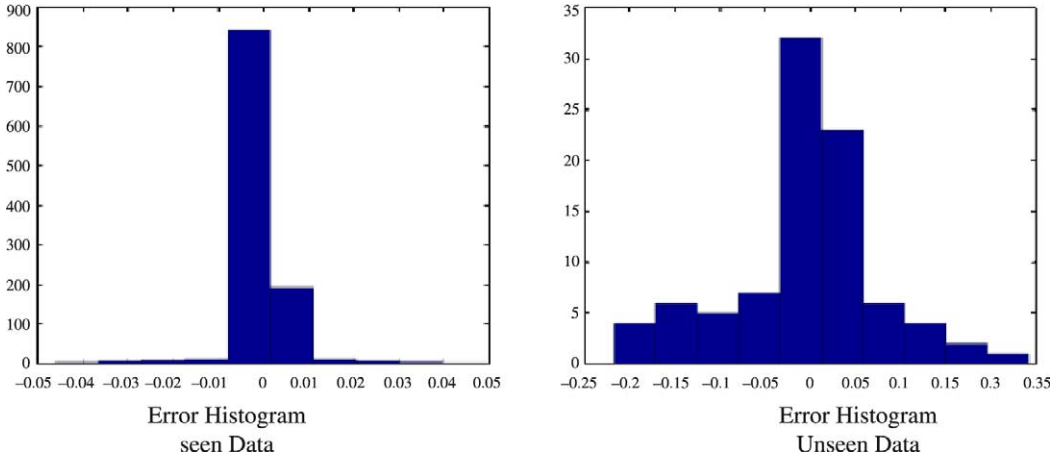

Fig. 13. Error of the mapping of the FETM using a RBFN.

with the RBFN. Using a similar technique Yangzhou and Xueyin [38] showed how a uniform function (i.e. a function that is not person specific) achieves results of $a_{\text{avg}} = 0.51$. This technique improves on this by computing a uniform function that achieves considerably better results. Fig. 13 shows the error of the mapping within the FETM. The histogram on the left is due to the mapping error for all images in the training set and the histogram on the right shows the error for all the unseen images.

6. Conclusion and future work

This paper demonstrated the construction of a universal mapping function that maps a neutral image of a face to one depicting a desired facial expression. This was achieved by constructing a FESM and a FETM. Using these models, several networks were trained which could accurately map a neutral image of a face to an image of the same subject portraying an alternative expression.

These models were based on the FACS, an anatomical analysis of facial actions. The FACS provided us with a universal method of analysing facial expression and allowed for the generation of shape and texture models that were independent of subject (age, sex, skin, colour, etc.).

A FHMN was used to compute mapping functions which map an image of a neutral face to one depicting a AU6 + AU12 + AU25, AU1 + AU2 + AU5 + AU26 and AU15 + AU17. This type of network achieved good results with the FESM but poor results with the FETM as Fig. 9 illustrates.

This network over generalised the mapping and hence much of the identity of a subject was lost during the calculations. To improve the results on both models a linear network was used with the FESM and a more sophisticated RBF network was used with the FETM. These networks greatly improved the results and a correlation coefficient between synthesised and authentic images on unseen data of $a_{\text{avg}} = 0.757$ was achieved. The results can be seen more clearly in Fig. 10. 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, age and skin colour.

References

[1] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis, Parameterized Facial Expression Synthesis Based on MPEG-4, EURASIP Journal on Applied Signal Proccessing 10 (2002) 1021–1038.

[2] Zhang, Q. Liu, Z. Guo, B. Shum, H. Geometric driven Photorealistic Facial Expression Synthesis, SIGGRAPH Symposium on Computer Animation, 2003.

[3] H. Wang, N. Ahujam, Face Expression Decomposition, Computer Vision and Robotics laboratory, Beckman Institute, UK, 2004.

[4] B. Choe, H.S. Ko, Analysis and synthesis of facial expression with hand generated muscle actuation basis Proceedings of computer animation, 2001 pp. 12-19.

[5] L. Gralewski, N. Campbell, B. Thomas, C. Dalton, D. Gibson, Statistical Synthesis of facial expressions for the protraying of emotion, International conference on computer science (2004) 190–203.

[6] S. Krinidis, I. Buciu, I. Pitas, Facial Expression Analysis and Synthesis: A survey Proceedings of the 10th International Conference on Human-Computer Interaction, 2003 pp. 1432-1436.

[7] I. King, H.T. Hou, Radial basis network for facial expression synthesis ICONIP'96, 1996.

[8] N. Arad, N. Dyn, D. Reisfeld, Y. Yeshurun, Image warping by Radial Basis Functions: Application to Facial Expressions, CVGIP: Graphical Models and Image Processing 56 (2) (1994) 161–172.

[9] B. Abboud, F. Davoine, Mo. Dang, Facial expression recognition and synthesis based on an appearance model Signal Processing: Image Communication', Elsevier, Amsterdam, 2004.

[10] J. Ghent, J. McDonald, 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, 2003.

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

[12] V. Bruce, A. Young, Understanding face recognition, British Journal of Psychology 77 (1986) 305–328.

[13] G. Rhodes, S. Brake, A. Atkinson, Whats lost in inverted faces?, Cognition 47 (1993) 25–57.

[14] P. Ekman, W.V. Friesen, Facial Action Coding System, in:. (Eds.), Human Interaction Laboratory, Dept. of Psychiatry, University of California Medical Centre, San Francisco, Consulting Psychologists Press, Inc.,, Human Interaction Laboratory, Dept. of Psychiatry, University of California Medical Centre, San Francisco, Consulting Psychologists Press, Inc. 577 College Avenue, Palo Alto, California, 1978,. 94306

[15] M. Perret, J.K. Hietanen, P. Oram, P. Benson, The effects of lighting conditions on response of cells selective to face views in the macaque temporal cortex, Experiment Brain Research 89 (1992) 157–171.

[16] T.F. Cootes, C.J. Taylor, Statistical Models of Appearance for Computer Vision Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, UK October 26th, 2001.

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

[18] C. Landis, Studies of emotional reactions: II. General behavior and facial expressions, Journal of Comparitive Psychology 4 (1924) 447–509.

[19] R.I. Birdwhistell, Kinesics and Context, University of Pennsylvania, Philadelphia, 1970.

[20] G. Young, T.G. Decarie, An ethology-based catalogue of facial/vocal behaviours in infancy, Archives of Psyhology 37 (264) (1941).

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

[22] A.L. Yuille, D.S. Cohen, P. Hallinan, Feature extraction from faces using deformable templates, Int. J. Comput. Vision 8 (1992) 99–112.

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

[24] A. Hill, C.J. Taylor, Model based image interpretation using genetic algorithms, Image Vision Computer 10 (1992) 295–300.

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

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

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

[28] D. Vernon, Machine Vision, Automated Visual Inspection and Robot Vision, Prentice Hall, Berlin, 1991.

[29] G.E. Hinton, C.K.I. Williams, M.D. Revow, Adaptive elastic models for hand-printed character recognition Advances in Neural Information Processing Systems, 1992 p. 4.

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

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

[32] J.C. Gower, Generalised Procrustes Analysis, Psychometrika 40 (1975) 33–50.

[33] J.C. Principe, N. Euliano, W.C. Lefebvre, Neural and Adaptive Systems, Wiley, New York, 2000.

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

[35] J. Moody, C. Darken, Fast learning in Networks of locall-tuned processing units, Neural Computation 1 (1989) 281–294.

[36] H. Kobayashi, F. Hara, Recognition of Six Facial Expressions and Their Strength by Neural Network IEEE International Workshop on Robot and Human Communication, 1992 pp. 381-386.

[37] Cohn J. Kanade, Cohn-Kanade AU-Coded Facial Expression Database, Pittsburgh University, PA USA, 1999.

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