Facial expression synthesis using a statistical model of appearance

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

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

ABSTRACT
This paper details a procedure for generating a mapping function which maps an image of a neutral face to one depicting a smile. This is achieved by the computation 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 facial expression called the *Facial Action Coding System* (FACS). The FEAM and the FETM allow for the generation of a subject independent 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.

KEY WORDS
Image Processing and Analysis, Image Synthesis, Facial Expression Shape Model (FESM), Facial Expression Texture Model (FETM)

1 Introduction

Facial expressions play a major role in how people communicate information. 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. The development of this mapping function involves a comprehensive understanding of expression. In the past facial expressions have been studied by cognitive psychologists [1, 2], social psychologists [3], neurophysiologists [4], computer scientists [5] and cognitive scientists [6].

The model of facial expression described in this paper is Ekman's [3] 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's in question from a number of subjects showing that expression in a training set. 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 Functions* (RBF) to generate subject independent mapping functions and present the results in this paper.

2 Measuring expression

Few studies have measured how the face moves as an expression forms [7, 8, 3, 9, 10]. 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.

According to Faigin [11], of the twenty-six muscles that move the face, only eleven 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 ex-

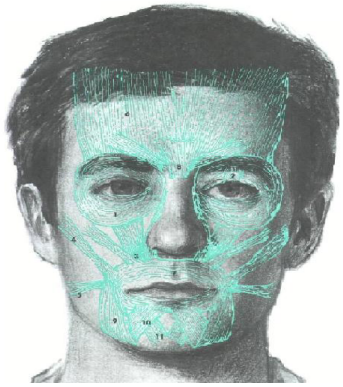

Figure 1. The eleven most influential muscles in expression formation, (Used by permission [11]).

pressions 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.

3 Computational shape and texture models

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. *Hand crafted models* can be developed using circles, lines, and arcs, that can move around relative to one another. Yuille *et al.* demonstrated this technique in modelling parts of the face such as the eyes and mouth [12]. Lipson *et al.* [13] and Hill *et al.* [14] 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 [15], and Grimson [16] demonstrate the effectiveness of this technique by locating an object within an image.

The two most common techniques for representing shapes are $active\ contour\ models$ or $snakes$ and the $Fourier\ series\ shape\ model$. $Active\ contour\ models$ or $snakes$ have been demonstrated to be very effective in this domain [17]. These energy minimizing curves are modelled as having stiffness and elasticity and are attracted toward features such as lines and edges. Staib and Duncan [18] use the $Fourier\ series\ shape\ model$ technique effectively to describe medical images and Bozma and Duncan [19] 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. 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. These points are weighted on their level of importance, depending on which AU the statistical model is being built for.

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 generalized Procrustes alignment (GPA). This technique aligns two shapes with respect to position, rotation and scale by minimizing 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 been mapped.

$Principal\ Component\ Analysis$ (PCA, also known as the Karhunen-Loéve transform) is a technique used to lower the dimensionality of a feature space. This method 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. For exact details of this process see [20, 23].

To calculate the *Facial Expression Texture Model* (FETM) we use a similar process. Firstly, we warp all images to the mean shape as described by the shape model. This is achieved using Delaunay triangulation to segment the mean shape into 214 separate regions using the landmark points as control points. 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 here again to analyse how the vectors move 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:

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

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

$$\delta \mathbf{x}_i^{k_j} = \mathbf{x}_i^{k_j} - \overline{\mathbf{x}} \tag{2}$$

where $\delta \mathbf{x}_i^{k_j}$ is the difference between $\mathbf{x}_i^{k_j}$ and the mean vector $\overline{\mathbf{x}}$. The covariance matrix is then calculated. In the experiments in this paper the n x n covariance matrix is very large, where $n = 65025$. For this reason 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 represented as

$$S = \frac{1}{s} D D^T \tag{3}$$

Let T be the s x s matrix

$$T = \frac{1}{s} D^T D \tag{4}$$

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 FESM [20, 23].

4 Function approximation

ANNs have proven to be successful in many practical problems. It has been shown that ANNs can recognise handwritten characters [24], spoken words [25] and more relevantly human faces [26]. 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.

A Feedforward Heteroassociative Memory Network (FHMN) can be used to compute a mapping from x to y [27]. 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 n principal components that represent a neutral face as input and the n principal components that represent a face depicting a specific expression as output. In this manner a mapping function is learned which maps the shape of a neutral face to that of a specific expression.

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 [27].

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 [28, 29]. 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

$$\hat{f}(x) = w_0 + \sum_{u=1}^{k} w_u \mathbf{G}_u(d(x_u, x)) \tag{5}$$

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.

5 Experiments and results

To create a FESM and a FETM of facial expression 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 [30]. 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 eighty 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 x 480 pixel arrays of greyscale values.

Each face was manually labelled using 122 landmark points and aligned with each other using Procrustes alignment. Following this each identity was aligned to itself using the landmark points that don't move as the expression forms. This feature alignment emphasizes the variance in shape as an expression forms and reduces the complexity of the mapping function. Fig 2 illustrates every shape aligned to the mean shape.

PCA was performed on the data and the top 15 principal components were used in the FESM. The top 15 principal components describe 96.59% of the total variance found in the training set. The mean shape was segmented

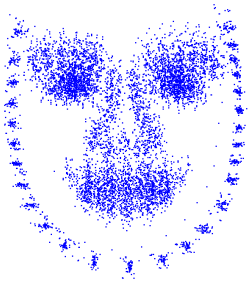

Figure 2. All shapes aligned to the mean shape

using Delaunay triangulation and each image was warped to the mean shape using a piece-wise affine transformation. The segmented mean shape can be seen in Fig 3.

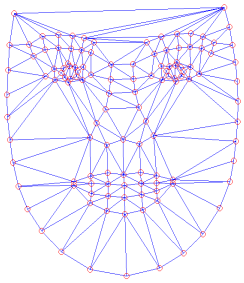

Figure 3. Mean shape segmented using Delaunay triangulation

The mean image was then calculated (Fig 4). 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.

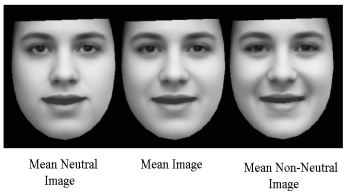

Figure 4. The mean images

A FHMN was used to generate a mapping from a neutral expression to one depicting AU 6, AU 12 and AU 25. This network was implemented on the shape and texture separately to create two mappings. Of the 40 subjects used to create the FESM and the FETM, 37 subjects were used during the training of network.

This network, using the FESM produced encouraging results but failed to return convincing results with the FETM. Fig 5 illustrates the effect of passing the shape and texture of an image through these mapping functions.

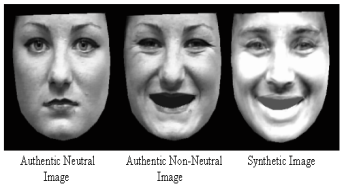

Figure 5. Expression Synthesis using a FHMN

To improve the mapping further we used a *Linear Network* (LN) with the FESM and implemented 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 subjects and 74 images. Three subjects were excluded from the training of the 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 FESM using a linear network and the FETM implementing a RBF network.

$Table_1$	$Experiment_1$			
$Subject$	$FESM$	LN	$FETM$	RBF
1	0.8770		0.9999	
2	0.7480		1	
3	0.4927		0.6017	
4	0.9575		0.6771	
5	0.9766		0.6208	
$Average$	0.8104		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 $t_{avg} = 0.6645$ and a correlation coefficient of $s_{avg} = 0.8089$ for the FESM. This results in an average of $a_{avg} = 0.7367$. Using a similar technique Yangzhou and Xueyin [31] showed how a $uniform\ function$ (i.e. a function that is not person specific) achieves results of $a_{avg} = 0.51$. This technique improves on this by computing a uniform function that achieves considerably better results. Fig 6 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 7 illustrates photo-realistic synthetic facial expression.

6 Conclusion and future work

This paper demonstrated the construction of a uniform 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

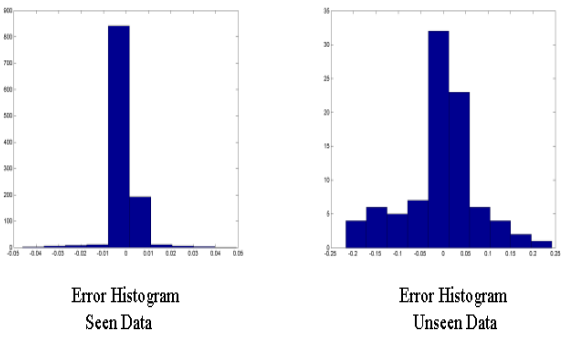

Figure 6. Error of the mapping

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 provides us with a universal method of analyzing facial expression and allowed for the generation of shape and texture models that were independent of subject (age, sex, skin colour etc.). The FESM and the FETM were build on top of this premise.

Each image was first labelled with 122 landmark points around the outline of face, the mouth, nose, pupil, eye and eyebrow. The images were aligned to each other using procrustes alignment and then each identity was aligned together using the landmark points that didn't move as an expression is formed. The mean shape was calculated and a difference vector was computed which consisted of the difference each shape had with the mean shape. PCA was performed on the data. The top 15 principal components of the FESM could explain 96.59% of the total variance found in the training set.

The mean image was segmented into 214 separate regions using Delaunay triangulation. Each image was warped to the mean shape using a piece-wise affine transformation. Every image was converted to a single vector and the difference vectors were calculated in the same manner as in the FESM. PCA was performed again on the data. 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 map an image of a neutral face to one depicting a smile (AU 6, AU 12, AU 25). This type of network achieved good results with the FESM but poor results with the FETM as Fig 5 illustrates. This network over generalized 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 synthesized and authentic images of $a_{avg} = 0.7367$ was achieved. The results can be seen more clearly in Fig 7. The first two rows of this diagram show expression synthesis on data that was

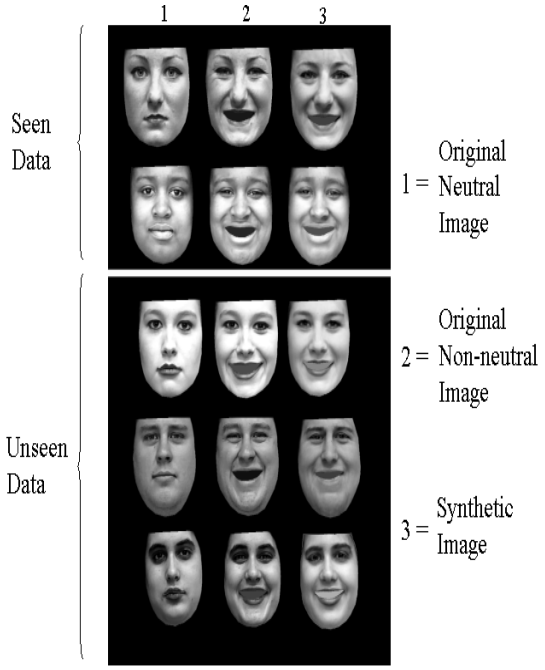

Figure 7. Real neutral, real non-neutral and synthesized images.

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 FESM and the FETM for expression classification. This could be done using similar neural networks to the ones detailed in this paper.

References

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

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

[3] 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.

[4] 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.

[5] 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.

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

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

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

[9] Birdwhistell, R.I “Kinesics and Context” Philadelphia: university of Pennsylvania Press, 1970.

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

[11] Faigan, G. “The Artist’s guide to Facial Expressions”, Watson-Guphill Publications, 1990.

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

[13] 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

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

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

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

[17] 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).

[18] 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

[19] 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

[20] 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

[21] 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

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

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

[24] 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.

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

[26] 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.

[27] Principe, J.C. Euliano, N.R and Lefebvre, W.C “Neural and Adaptive Systems”,Wiley, 2000

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

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

[30] Cohn, J. Kanade “Cohn-Kanade AU-Coded Facial Expression Database”, Pittsburgh University, 1999.

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