An Automatic Region Based Methodology for Facial Expression Recognition

Anastasios Koutlas, IEEE Student Member
Dept. of Medical Physics, Medical School
Unit of Medical Technology and Intelligent Information Systems
University of Ioannina
Ioannina, Greece
akoutlas@cc.uoi.gr

Dimitrios I. Fotiadis, IEEE Senior Member
Dept. of Computer Science
Unit of Medical Technology and Intelligent Information Systems
University of Ioannina
Ioannina, Greece
fotiadis@cs.uoi.gr

***Abstract*— This work investigates the use of a Point Distribution Model to detect prominent features in a face (eyes, brows, mouth, etc) and the subsequent facial feature extraction and facial expression classification into seven categories (anger, fear, surprise, happiness, disgust, neutral and sadness). A multi-scale and multi-orientation Gabor filter bank, designed in such a way so as to avoid redundant information, is used to extract facial features at selected locations of the prominent features of a face (fiducial points). A region based approach is employed at the location of the fiducial points using different region sizes to allow some degree of flexibility and avoid artefacts due to incorrect automatic discovery of these points. A feed forward back propagation Artificial Neural Network is employed to classify the extracted feature vectors. The methodology is evaluated by forming 7 different regions and the feature vector is extracted at the location of 20 fiducial points.**

***Keywords*— Active Shape Models, Gabor Filters, Facial Expression Recognition**

I. INTRODUCTION

Facial expression recognition determines the state of an expression in a human by a machine. The task to automatically analyse facial expressions by computing systems can be beneficial to many different scientific subjects such as psychology, neurology, psychiatry as well as applications for everyday life such as driver monitoring systems, automated tutoring systems or smart environments and human-computer interaction. Recognizing facial expressions automatically is a difficult task due to the non-uniform nature of the human face or limitations such as lightening conditions.

In 1971 Ekman et al. determined 6 basic emotions; anger, fear, surprise, happiness, disgust and sadness [1]. Basic emotions are universal and exist in different human ethnicities and cultures. The neutral face expression inherits the universality of the basic emotions and is usually included as a seventh basic expression. Even though the term emotion is used for categorization, emotions do not rely solely on visual information [2].

There are three main steps that an automatic facial expression recognition system should carry out. First, the face or facial prominent characteristics such as eyes, mouth, nose and so on, must be located; these will be processed. Next the feature extraction process will extract feature vectors to represent the image. There are two options to construct the feature vector; treating the face as a whole and extract the feature vector without taking into account prominent features of the face. Alternatively at the discovered locations of the facial prominent characteristics the feature vector is extracted usually using a mathematical transformation so that the facial properties of the face are represented in an appropriate manner. At the end the extracted feature vectors are classified into the appropriate expressions.

There exist several approaches to detect a face from single images or image sequences, these include knowledge based methods, template based methods or appearance based methods [3], [4]. Template based methods are simple to implement but are usually prone to failure when large variations in pose or scale exist [3]. In part the above problem can be tackled by deformable models. Kass et al. used the Active Contour Models or snakes [5]. The snake is initialized at the proximity of the structure and is fitted onto nearby edges. The evolution of the snake relies on the minimization of an energy function. Cootes et al. has used Active Shape Models (ASM) [6] and Active Appearance Models (AAM) [7]. Active Shape Models differ from snakes mainly due to global shape constrains that are enforced at the deformable model, ensuring this way that the model deforms into shapes that are allowed by the global shape constrains. Moreover a statistical gray-level model is built around landmark points which assume a Gaussian and unimodal distribution. Active Appearance Models extend the functionality of ASM capturing texturing information along with shape information. Recently variations of the ASM method have been introduced. Optimal Features ASM (OF-ASM) [8] allows for multimodal distribution of the intensities while high segmentation accuracy is reported but is more computationally expensive. Sukno et al. [9] extended OF-ASM to allow application in more complex geometries using Cartesian differential invariants. Methods similar to ASM employing a point distribution model to fit the shapes, expand into 3-dimensional problems [10], [11].

When the face is located and its prominent features identified, a mathematical representation is used so that feature

vectors are extracted. Two approaches are used to represent the face and consequently the facial features. The first, often referred to as holistic approach, treats the face as a whole. Essa and Petland [12] treated the face holistically using optical flow and measured deformations based on the face anatomy. Donato et al. [13] has used several methods for facial expression recognition. Fisher linear discriminates (FLD) were used to project the images in a space that provided the maximal separability between classes and Independent Component Analysis (ICA) to preserve higher order information.

The other approach referred to as local approach, tries to symbolize the geometry of prominent features in a local manner. Fiducial points are used around the prominent features of the face, the location of which are used to extract the feature vector. The number of fiducial points used varies and mainly depends on the desired representation, as it is reported that different positions hold different information regarding the expressions [14]. The way that these fiducial points are identified in an image can either be automatic [15] or manual [14], [16]-[17].

It has been shown that simple cells in the primary visual cortex can be modeled by Gabor functions [18]-[19]. This solid physiological connection between Gabor functions and human vision has yielded several approaches to feature extraction [20] and facial expression recognition [14]-[17],[21]-[23]. Zhang et al. [17] compared the Gabor function coefficients with the coordinate positions of the fiducial points and concluded that the first represent the face better than the latter. Donato et al. [13] reported that Gabor functions performed better than any other method used in both analytic and holistic approaches.

In this work an automatic approach to facial expression recognition is presented based on features extracted at the location of fiducial points using Gabor filters. The methodology is based on automatically locating 74 landmark points using Active Shape Models (ASM). ASMs were chosen for their ability to locate specific landmark points, accurately, with a low computational footprint. The discovery of the location of the landmark points and subsequently the fiducial points is completely automatic. A set of 20 fiducial points, which is derived from the 74 landmarks, is proposed for the feature extraction process, around prominent features of the face that contain the most significant information regarding the muscle movement which is responsible for facial expressions. The proposed approach forms the feature vector from a region around each fiducial point utilizing more information in the feature vector. Furthermore, artefacts are avoided since certain fiducial points are identified less precisely than when identified by a human. The methodology is based on the processing of an image by a Gabor filter bank at the selected locations, which is specially designed to avoid redundant information. The methodology is evaluated using the Japanese Female Facial Expression (JAFFE) database [22]. The accuracy of the methodology is compared with the corresponding one using 34 points manually located in the subjects face. The reduced set of fiducial points can perform as good as methods with more points proposed in the literature but also the dimensionality of the feature vector is decreased, leading to a cost efficient method.

II. MATERIALS AND METHODS

The proposed methodology includes four stages: (a) automatic discovery of prominent features of a face, such as the eyes, and the discovery of fiducial points, (b) construction of the Gabor Filter Bank, (c) extraction of the Feature vector at the location of the fiducial points and (d) classification (Figure 1).

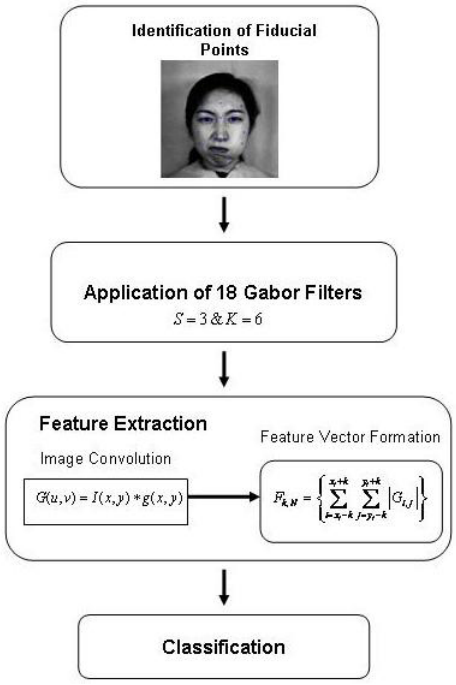

Figure 1. Flow chart of the proposed methodology

A. Active Shape Models

Active Shape Models [6] utilize information from points around prominent features of the face which are called landmarks. A Point Distribution Model (PDM) and an image intensity profile are computed around the landmarks.

For a total of s landmark points the coordinates are represented in a single vector or shape as

$$\mathbf{x} = (x_1, \cdots, x_s, y_1, \ldots, y_s)^T. \quad (1)$$

The shapes collected from the training stage are aligned to the same coordinate frame. The dimensionality of the aligned data is reduced by applying Principal Component Analysis and the mean shape is computed, thus forming the PDM. Any shape of the training set can be approximated by the mean shape, $\bar{\mathbf{x}}$, the eigenvector matrix $\mathbf{P}$ and b_i, which defines the shape parameters for the i^{th} shape,

$$\mathbf{x}_i = \bar{\mathbf{x}} + \mathbf{P}b_i, \quad b_i = \mathbf{P}^T(\mathbf{x}_i - \bar{\mathbf{x}}). \quad (2)$$

The dimensionality is reduced by selecting only the eigenvectors that correspond to the largest eigenvalues. Depending on the number of excluded eigenvectors there is an

error introduced in (2). Furthermore, the parameter b_i is constrained to deform in ways that are found in the training set:

$$|b_i| \le \beta\sqrt{\lambda_i}, \quad 1 < i < M, \tag{3}$$

where β is a constant, usually, from 1-3, λ_i is the i^{th} eigenvalue and M is the total number of the selected eigenvectors. This is done to ensure that only allowable shapes are represented by (2).

At the training stage, for each point a profile that is perpendicular to the shape boundary is investigated to obtain information regarding the gray-level structure above and below each point. A vector is computed using the intensity derivatives along the profile. This is done to ensure some tolerance to global intensity changes. Each sample is then normalized using the statistical model gathered from all training images for that point. Under the assumption that the samples are part of a Gaussian distribution the mean and the covariance are calculated. The above procedure is repeated for all landmark points thus forming a statistical gray-level structure model. The correct deformation and convergence of a shape in a new image is done recursively. First, the mean shape is initialized. The goal is to deform each point of the shape so that its correct position is located. In order to identify the correct position for any given point a profile perpendicular to the shape model is investigated. This is the same procedure as in the training stage. The displacement for each landmark point is estimated by minimizing the Mahalanobis distance between the training model and the test model. The shape parameters are updated and the procedure is repeated until the point converges to a correct location. This procedure is repeated for all points until convergence to correct locations.

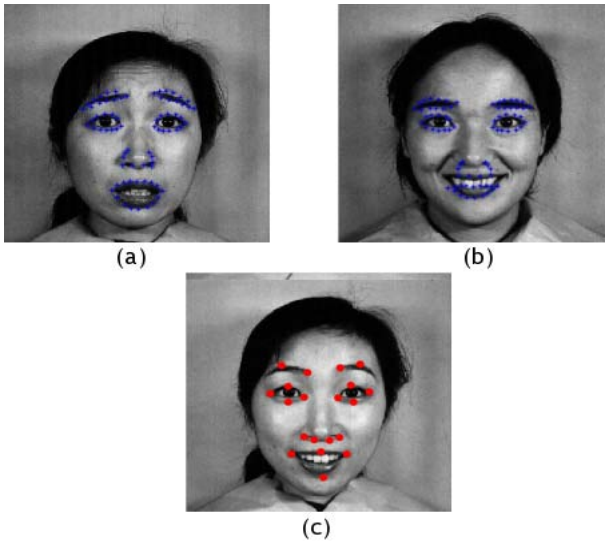

Figure 2. (a) Correctly identified, (b) incorrectly identified and (c) the 20 proposed points

For each image a total number of 74 points are chosen to locate the landmark points. The number of fiducial points that are used in the feature extraction process is reduced to 20. The points that are chosen are near the places of interest in the face that contain information about the muscle movement. Figure 2 shows two examples of images that the prominent features were (a) correctly identified and (b) incorrectly identified and the set of the 20 fiducial points proposed for the feature vector extraction.

B. Gabor Function

A two dimensional Gabor function $g(x, y)$ is the product of a 2-D Gaussian-shaped function referred to as the envelop function and a complex exponential (sinusoidal) known as the carrier and can be written as [18]-[19], [23]:

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi jW\right], \tag{4}$$

where x, y are the image coordinates, σ_x, σ_y are the variances in the x, y coordinates respectively and W is the frequency of the sine wave. The above representation combines the even and odd Gabor functions which are defined in [18].

Its Fourier Transform $G(u, v)$ can be written as,

$$G(u,v) = \exp\left\{-\frac{1}{2}\left[\frac{(u-W)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}\right]\right\}, \tag{5}$$

where $\sigma_u = 1/2\pi\sigma_x$ and $\sigma_v = 1/2\pi\sigma_y$.

C. Gabor Filter Bank

A Gabor filter bank can be defined as a series of Gabor filters at various scales and orientations. The application of each filter on an image produces a response for each pixel with different spatial-frequency properties.

Let $g(x, y)$ be the mother function, the Filter bank derives by scaling and rotating the mother function:

$$g'(x,y) = g(x',y'), \quad \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}, \tag{6}$$

where $\theta = n\pi / K$, K is the total number of orientations and $n = 0, 1, \ldots, K-1$.

Manjunathan showed that Gabor filters include redundant information in the images produced by the filter [23], [16]. By selecting certain scaling parameters the constructed filters are not overlapping with each other thus avoiding redundant information. This leads to the following equations for the filter parameters a, σ_u and σ_v:

$$a = \left(\frac{U_h}{U_l}\right)^{\frac{1}{S-1}}, \quad W = a^m U_l, \tag{7}$$

$$\sigma_u = \frac{(a-1)W}{(a+1)\sqrt{2\ln 2}}, \quad (8)$$

$$\sigma_v = \tan\left(\pi/2K\right)\sqrt{\frac{W^2}{2\ln 2} - \sigma_u}, \quad (9)$$

where a is the scaling factor, S is the number of scales, $m = 0, 1, \ldots, S-1$, and U_h and U_l are the high and low frequencies of interest.

In this work $U_h = \sqrt{2}/4$, $U_l = \sqrt{2}/16$ are chosen with three scales ($S = 3$) and six orientations ($K = 6$) differing by $\pi/6$. Thus, 18 complex Gabor filters are defined in total which will be used to extract the feature vector for each image.

D. Feature Extraction

For any given image its Gabor decomposition at any given scale and orientation can be obtained by convolving the image with the particular Gabor filter.

$$G(u,v) = I(x,y) * g(x,y) \quad (10)$$

The magnitude of the resulting complex image is given

$$|G| = \sqrt{Re(G)^2 + Im(G)^2}. \quad (11)$$

All features derive from $|G|$ and the feature vector $F_{k,N}$ is formed:

$$F_{k,l} = \left\{ \sum_{i=x_l-k}^{x_l+k} \sum_{j=y_l-k}^{y_l+k} |G_{i,j}| \right\}, \; l = 0,1,..,N, \; k = 0,1,..,5, \quad (12)$$

where N is the number of the fiducial points, a total number of 20 points are used to form the feature vector. k is the number of neighboring pixels used to form the regions.

E. Artificial Neural Networks

A feed forward back propagation ANN is employed. The architecture of the ANNs consists of three layers. The first layer (input layer) consist of t input nodes where t is the dimension of the feature vector ($F_{k,N} \in R^t$). The second layer (hidden layer) consists of $(t+c)/2$ neurons, where c is the number of the classes. The sigmoid function is used as activation function for these hidden neurons. Finally the third layer (output layer) consists of c neurons. The activation function of the output neurons is the linear function. In order to train the ANN the mean square error function is used and the number of epochs is 500.

F. Dataset

The Japanese Female Facial Expression Database (JAFFE) [22] database is used for the evaluation of the proposed method. It features ten different Japanese women posing 3 or 4 examples for each basic emotion containing a total of 213 images. Neutral position inherits all characteristics of a basic emotion and is included in the annotation of the database as a seventh basic emotion.

III. RESULTS

Seven sets of experiments have been conducted with the automatic identification of fiducial points and are compared with seven sets of experiments conducted when 34 fiducial points are manually identified. Table I presents the accuracy of the methodology for both sets of points and all different regions that are used. In Table II the abbreviations correspond to the 7 categories that are used for the classification (SU for surprise, DI for disgust, FE for fear, HA for happy, NE for neutral, SA for sadness and finally AN for anger). For the evaluation the ten fold stratified cross validation method is used. The gradual increase points out the when the region gets broader utilizes more information that describe better facial geometry. It should be noted that the dimension of the feature vector when the 20 points are used is 360 while when 34 points are used the dimension is 612.

TABLE I. ACCURACY OBTAINED FOR DIFFERENT REGION SIZES

Region size	Accuracy	
	Automatic 20 points	*Manual 34 points*
Single Pixel	67.6%	72.8%
3x3	77.0%	81.7%
5x5	84.0%	84.0%
7x7	83.1%	85.0%
9x9	90.2%	87.3%
11x11	89.7%	87.8%
13x13	87.3%	87.0%

TABLE II. CONFUSION MATRIX OF THE BEST PERFORMING REGION (9X9) FOR THE 20 POINTS SET

	SU	DI	FE	HA	NE	SA	AN
SU	28	0	1	0	1	0	0
DI	0	26	2	0	0	1	0
FE	1	2	26	0	1	2	0
HA	0	0	1	29	1	0	0
NE	0	0	0	0	30	0	0
SA	0	1	4	1	0	25	0
AN	0	1	0	0	0	0	28

The best accuracy is reported when a region of 9x9 pixels is used for the 20 fiducial points set. In Table II the confusion matrix of the best performing region is presented. Fear and sadness have the poorest performance amongst all emotions while neutral has the highest. There are a few misclassifications of sadness that are classified as fear. Zhang et al. [17] have

excluded fear from their experiments due to the difficulty to express the emotion from the subjects and some evidence that fear is processed differently by the human brain.

IV. DISCUSSION

An automated facial expression recognition methodology is presented. The identification of the prominent features is done automatically and the feature vector is extracted using a specially constructed Gabor Filter bank that avoids redundant information and a region based methodology which ensures some flexibility on the identified points and avoids artefacts. Moreover a 20 fiducial point set is proposed that models facial geometry adequately for facial expression recognition.

Zhang et al. [17] have performed a set of experiments extracting the feature vector by single pixels at the location of 34 fiducial points manually identified and a modified ANN. When they used the full annotation of JAFFE they reported less than 90% accuracy. They repeated the experiments excluding fear and reported accuracy of 92.3%. Guo and Dryer [16] compared the performance of different classifiers on the JAFFE database using 34 fiducial points manually identified. They extracted the feature vector using the magnitude of the pixel values of the 34 fiducial points. When used the Simplified Bayes they reported accuracy of 63.3%, when used linear Support Vector Machines (SVM) 91.4% and when used non linear (Gaussian Radial Basis function kernel) SVM 92.3%. The performance of SVM suggests that are suitable for this problem and it is aimed to investigate in the future how they apply to the region based approach. The methodology presented here utilizes 20 fiducial points that are discovered automatically. The set of fiducial points decreases the dimensionality of the feature vector ~40% without any loss in terms of accuracy. The pixel-based approach was modified to accommodate information from neighboring pixels, called regions. This is done to ensure that artifacts are avoided due to imprecise identification of prominent features of the face when done automatically which differs than when identified by a human manually. The methodology presented has an accuracy of 90.2% but does not perform very well when trying to classify sadness or fear and reports the biggest losses between the two emotions. Difficulties in distinguishing certain emotional states (sadness-fear and disgust-anger) even among human experts are reported by Yin et al. [24].

Further improvement of the methodology includes the use of sequential images and the use of a three dimensional filter bank, including time as a parameter along with the use of a different classifier that presents increased performance.

ACKNOWLEDGMENT

This work was partly funded by the General Secretariat for Research and Technology of the Hellenic Ministry of Development (PENED 2003 03OD139).

REFERENCES

[1] P. Ekman and W.V. Friesen, "Constants across cultures in the face and emotion," J. Pers. Soc. Psychol., vol 17 (2), pp. 124-129, 1971.

[2] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," Pattern Recognition, no. 36, pp. 259-275, 2003.

[3] M.H. Yang, D.J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," IEEE Trans on Pattern Analyis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, 2002

[4] E. Hjelmas, "Face Detection: a survey," Computer and Image Understanding, vol. 83, pp. 236-274, 2001

[5] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour Models," Proc. First IEEE Int'l Conf. Computer Vision, pp. 259-269, 1987

[6] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active shape models – Their training and application," Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38-59, 1995

[7] T.F. Cootes, G. Edwards, and C.J. Taylor, " Active appearance models," Proc. European Conf. Computer Vision, vol. 2, pp. 484-498, 1998

[8] B. van Ginneken, A.F. Frangi, J.J. Staal, B.M. ter Har Romeny, and M.A. Viergever, "Active shape model segmentation with optimal features," IEEE Trans. Medical Imaging, vol. 21, no. 8, pp. 924-933, 2002.

[9] F.M. Sunko, S. Ordaas, C. Butakoff, S. Cruz, and A.F. Frangi, "Active shape models with invariant optimal features: Application to Facial Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 7, pp. 1105-1117, 2007

[10] P. Nair, A. Cavallaro, "Region Segmentation and Feature Point Extraction on 3D Faces Using a Point Distribution Model," IEEE Int. Conf. Image Processing, pp. 85-88, 2007.

[11] T.J. Hutton, B.R. Buxton, P. Hammond, "Dense surface point distribution models of the human face," MMBIA 2001, pp. 153-160, 2001.

[12] I. Essa, A. Pentland, "Coding, Analysis, Interpretation, Recognition of Facial Expressions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19 , no. 7, pp. 757-763, Jul. 1997.

[13] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, T.J. Sejnowski, "Classifying Facial Actions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 974-989, Oct. 1999.

[14] M.J. Lyons, J. Budynek, S. Akamatsu, "Automatic Classification of Single Facial Images," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 12, pp. 1357-1362, Oct. 1999.

[15] H. Gu, Y. Zhang, Q. Ji, "Task oriented facial behavior recognition with selective sensing," Computer Vision and Image Understanding, vol. 100, pp. 385-415, 2005.

[16] G. Guo, C.R. Dyer, "Learning From Examples in the Small Sample Case: Face Expression Recognition," IEEE Trans. Sys. Man and Cybernetics–PART B: Cybernetics, vol. 35, no. 3, pp. 477-488. 2005.

[17] Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu, "Comparison Between Geometry-Based and Gabor-Wavelet-Based Facial Expression Recognition Using Multi-Layer Perceptron," Proc. 3rd Int. Conf. Automatic Face and Gesture Recognition, pp. 454-459, 1998.

[18] J. Dougman, "Two-dimensional spectral analysis of cortical receptive field profiles," Vision Research, vol. 20, pp. 846-856, 1980.

[19] J. Dougman, "Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical fields," J. Opt. Soc. Am. A., vol. 2, no. 7, Jul 1985.

[20] J. Ye, Y. Zhan, and S. Song, "Facial expression features extraction based on Gabor wavelet transformation," IEEE Int. Conf. Systems, Man and Cybernetics, pp. 2215-2219, 2004

[21] W. Liu, Z. Wang, "Facial Expression Recognition Based on Fusion of Multiple Gabor Features," Proc. 18th Int.Conf. on Pat. Rec., 2006.

[22] M. Lyons, S. Akamatsu, "Coding Facial Expressions with Gabor Wavelets," Proc. 3rd Int. Conf. Automatic Face and Gesture Recognition, pp. 200-205, 1998.

[23] B.S. Manjunath, W.Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 837-842, Aug. 1996.

[24] L. Yin, X. Wei, Y. Sun, J. Wang and M. Rosato, "A 3D facial expression database for facial behavior research," IEEE Int'l Conf. Face and Gesture Recognition, pp. 211-216, 2006.