Image Processing and Machine Learning Techniques for Facial Expression Recognition

Anastasios Koutlas
Medical Physics Department, Medical School, University of Ioannina,GR 45110, Ioannina, Greece
Dimitrios I. Fotiadis
Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, GR 45110, Ioannina, Greece

ABSTRACT

The aim of this chapter is to analyse the recent advances in image processing and machine learning techniques with respect to facial expression recognition. A comprehensive review of recently proposed methods is provided along with an analysis of the advantages and the shortcomings of existing systems. Moreover, an example for the automatic identification of basic emotions is presented; Active Shape Models are used to identify prominent features of the face; Gabor filters are used to represent facial geometry at selected locations of fiducial points and Artificial Neural Networks are used for the classification into the basic emotions (anger, surprise, fear, happiness, sadness, disgust, neutral). Finally, the future trends towards automatic facial expression recognition are described.

INTRODUCTION

The face is the fundamental part of day to day interpersonal communication. Humans use the face along with facial expressions to denote consciously their emotional states (anger, surprise, stress, etc.) or subconsciously (yawn, lip biting), to accompany and enhance the meaning of their thoughts (wink) or exchange thoughts without talking (head nodes, look exchanges). Facial expressions are the result of the deformation in a human's face due to muscle movement. The importance of automating the task to analyse facial expressions using computing systems is apparent and can be beneficial to many different scientific subjects such as psychology, neurology, psychiatry, as well as, applications of everyday life such as driver monitoring systems, automated tutoring systems or smart environments and human-computer interaction. Although humans are able to identify changes in facial expressions easily and effortlessly even in complicated scenes, the same is not an easy task to be undertaken by a machine. Moreover, computing systems must share the same robustness and accuracy with a human so that these systems could be used in a real-world scenario and provide adequate aid.

Advances in topics such as face detection, face tracking and recognition, psychological studies as well as the processing power of modern computer systems make the automatic analysis of

facial expressions possible for use with real world examples where responsiveness (i.e. real time processing) is required along with sensitivity (i.e. being able to detect various day to day emotional states and visual cues) and the ability to tolerate head movements or sudden changes.

For an effective automatic facial expression recognition (AFER) system there are several characteristics that must be present so that it can be efficient. These are outlined in the Figure 1.

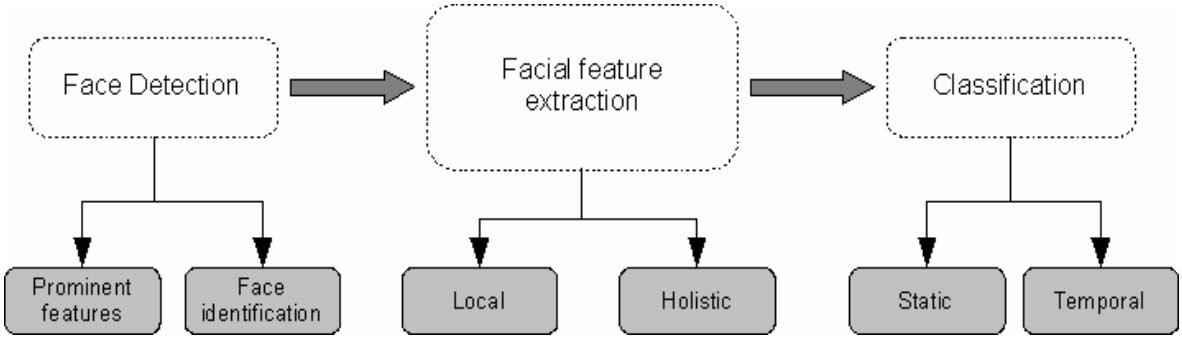

Figure 1 Structure of an automatic facial expression recognition system

Face detection and identification of prominent features is a crucial step for an AFER system. It is the first step for any system that carries the automatic tag and the performance of this step in terms of accuracy is crucial for the overall accuracy of the system. Various approaches are presented in the literature in terms of static or temporal identification of the face or identification of prominent features such as eyes in contrast to identifying the presence of a face in a scene.

When the face is located it must be modeled so that it can be represented in an appropriate manner. The facial representation could be based on the facial geometry that encompasses some unique features of homogeneity and diversion across humans. It could also be based in characteristics that appear after some transformation with mathematical expressions modeling texture, position and gray-level information. After that the feature vector is built by extracting features. It can be represented either holistically or locally. Holistic approach treats the face as a whole, i.e. the processing of the face and the mathematical information applies to the whole face without considering any special prominent features of it. On the other hand the local approach treats each prominent feature of the face in a different way and the feature extraction process is applied in selected locations in the image which are often called fiducial points. Lastly, there are systems which are related to the processing of image sequences or static images which combine the two approaches, treating the face in a hybrid manner. There is also a distinction in terms of the presence of temporal information or not.

Classification is the last step for an AFER system. The facial actions or the deformations due to facial movement are categorized either as basic emotions or as Action Units (AUs). In what follows depending on the use of temporal characteristics or not the classification process is considered temporal or static for this chapter.

This chapter introduces recent advances in automatic facial expression recognition. The first part contains an introduction to the automatic facial expression recognition systems, including their structure, their objectives and their limitations. In the second part a review of recent work, is presented related to face identification, acquisition and recognition, facial feature transformation, feature vector extraction and classification. In part three a particular approach is described along with quantitative results.

BACKGROUND

Introduction

Most systems try to recognize a small set of prototypic emotions which share characteristics of universality and uniformity across people with different ethnic background or cultural heritage. The six basic emotions were proposed by Ekman and Friesen (1971) and are: disgust, fear, joy, surprise, sadness and anger. The neutral position inherits most of the characteristics that are shared across basic emotions and could be considered a seventh basic expression. Diversity of the neutral position arises mainly due to variations in pose and not muscle movement.

In every day life basic emotions occur rather infrequently. Emotions that are more frequent to occur in everyday life are due to subtle changes in certain specific areas such as the eyebrows or eyelids and so on. For example the tightening of the lips in case of anger or the lowering the lips in case of sadness. These changes in the appearance of facial expression are subtle and systems that recognize such changes are required to be more precise. The Facial Action Coding System (FACS) (Ekman & Friesen, 1978) provides the mechanisms to detect facial movement by human coders. When a coder is viewing a sequence of the facial behaviour of a human subject can decode Action Units (AU). Action Units are a set of actions that correspond either to muscle movement in facial expressions such as raising upper lip or blinking or some miscellaneous actions such as bite lip or blow. FACS consists of 44 action units. There is also a scale of intensity that can describe each action unit in a scale of 5.

Even though Ekman and Friesen proposed certain combinations of action units as descriptive of certain emotions, FACS itself does not contain any emotion-specific information. These are coded in separate systems such as the Emotional Specific FACS (EMFACS) (Friesen & Ekman, 1983). By converting action units from FACS to EMFACS or other emotion-specific systems, expressions can be coded, such as sadness or surprise.

It is reported in the literature that there is a distinction between facial expressions that are spontaneous and those that are initiated by request often referred to as posed (Ekman, 1991,2003). From a physiological point of view it is perfectly justified since spontaneous actions and posed actions originate from different parts of the brain; namely the subcortical areas of the brain and the cortical motor strip, respectively (Meihle, 1973). Major differences between spontaneous and posed facial expressions are the actual movement that is initiated from facial muscles and the dynamics of the expression (Ekman & Rosenberg, 2005). Subcorticaly initiated facial expressions (spontaneous) are characterized by synchronized, smooth, symmetrical, consistent and reflex-like facial muscle movement. On the other hand facial expressions that are cortically initiated (posed) tend to be less smooth, with more varying dynamics (Ekman & Rosenberg, 2005).

To develop and evaluate systems that are subject to the above conditions reliable annotated databases must be used. There are several attempts in the literature for the development of such databases but it is difficult to comprehend all different variability issues in a single database. An example is the Japanese Female Facial Expression Database (JAFFE) (Lyons et al. 1999). It features ten different Japanese women posing 3 or 4 examples for each basic emotion containing a total of 213 still images. The Cohn-Kanade database (Kanade et al. 2000) is another database but differs from JAFFE since it contains temporal information and is used widely for facial expression analysis (Tian et al. 2001). It contains image sequences of 100 subjects posing a set of 23 facial displays and contains FACS annotation in addition to basic emotion tags. Although it is

used widely for the evaluation of AFER systems it has certain drawbacks. The image sequences in order to be complete and fully functional should contain 3 states for the dynamics of each expression; the onset which is the initialization of the expression, the apex which is the peak of the expression and the offset where the expression declines. Unfortunately, the Cohn-Kanade database contains information that excludes the offset of the expression. Another shortcoming of the Cohn-Kanade database is that the images contain a timestamp that is overlapping with the subject's expression certain times. The MMI database (Pantic et al., 2005) contains both posed and spontaneous facial actions. Furthermore, it contains over 4000 videos as well as 600 static images. The images are coded based on FACS, either single action units or combinations, and basic emotions. Furthermore apart from frontal views, profile views are included. Another recently developed database is the Yin Facial Expression Database (Yin et al. 2006) which contains 3D facial expression information. The expression data includes 3D models, texture information and raw model data. It also provides a landmark point set for evaluating facial features segmentation techniques. It also features 6 basic emotions plus the neutral position.

Most research groups that are working with AFER systems either use the available databases or collect their own signals to evaluate the methods. This slight fragmentation on the evaluation of such systems does not make possible the comperative evaluation of all methods proposed in the literature.

Face Detection

Face detection and identification of prominent features is a crucial step for an AFER system. This is the first step of any system that operates automatically and the overall performance of the system mainly depends on the correct identification of the face or certain facial features such as eyes, eyebrows, mouth and so on. The task of locating the face or the prominent features of a face in a scene should be independent of any occlusions in the scene, variations of lightening conditions and should tolerate changes in face pose. There are various approaches to detect faces or prominent characteristics of the face using appearance based methods and statistical techniques, or template based methods (Hjelmas & Low 2001; Yang et al. 2002; Li & Jain 2005).

The most commonly employed face detection algorithm in automatic facial expression recognition systems is the real-time face detector proposed by Viola and Jones (2001,2004). The face detector does not work directly with image intensities but there is a set of features extracted related to Haar basis functions. The Haar-like features can be computed at different scales and locations. For each set of features Adaboost is used to choose the most important features from the large set of potential features (Freund & Schapire, 1995). The classifiers are combined in a cascade, successive manner to speed up the detector's performance. The face detector is able to detect faces very rapidly. There are other works that have adapted the proposed methodology. Fasel et al. (2005) used Gentleboost (Friedman et al., 2000) instead of Adaboost. Gentleboost instead of using the binary output of each filter, uses the output in a continuous manner.

Statistical learning techniques combined with appearance features are usually used to detect faces in images. Rowley et al. (1998) used a neural network to detect face regions from non face regions using as a feature vector pixel intensities and spatial relationships between pixels. Sung and Poggio (1998) used a neural network also but as the feature vector they have used distance measures. A real-time method proposed by Petland et al. (1994) detects faces by using a view-based eigenspace method that incorporates prominent features of the face such as eyes and mouth. Apart from real-time processing the method can handle head positions which vary. Another method that can handle varying head motion was proposed by Schneiderman and

Kanade (2000) which utilises a 3D object detection and appearance features such as object or non object features using a product of histograms which contains object statistics based on wavelets coefficients and their position on the object.

Template based methods are simple to implement but are usually prone to failure when large variations in pose or scale exist (Yang et al., 2002). In part the above problem can be tackled by deformable models. Kass et al. proposed the Active Contour Models or snakes (Kass et al., 1987). The snake is initialized at the proximity of the structure and is fitted onto nearby edges. The evolution of the snake relies in the minimization of an energy function. Cootes et al. has proposed Active Shape Models (ASM) (Cootes et al. 1995) and Active Appearance Models (AAM) (Cootes et al. 1998). Active Shape Models differ from snakes mainly due to global shape constrains that are enforced on the deformable model, ensuring this way that the model deforms according to the variations of the landmark points found in the training set. 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) (Van Ginneken et al. 2002) allow for multimodal distribution of the intensities while high segmentation accuracy is reported but it is more computationally expensive. Sukno et al. (2007) extended OF-ASM to allow application in more complex geometries using Cartesian differential invariants.

Readers are referred to Hjelmas and Low (2001), Yang et al. (2002) and Li & Jain (2005) for a more thorough analysis concerning developments in detecting faces in images or image sequences.

Facial Features Extraction

The facial feature extraction step aims at modeling the face using some mathematical representation in such a way so that it could later form the feature vector and be fed into a classifier. There are two approaches to represent the face and subsequently facial geometry. Firstly, the face can be processed as a whole often referred to as holistic or analytic approach and secondly it can be represented at the location of specific regions or at the location of fiducial points often referred to as local approach.

Essa and Petland (1997) treated the face holistically using optical flow and measured deformations based on the face anatomy. Black and Yacoob (1998) also utilized an optical flow model of image motion for facial expression analysis. Their work explores the use of local parameterized optical flow models for the recognition of the six basic emotional expressions. Donato et al. (1999) has used several methods for facial expression recognition. They have used holistic Principal Component Analysis, EigenActions, where the principal components were obtained on the dataset by using difference images. A set of topographic local kernels were used for Local Feature Analysis that were matched to the second-order statistics of the input ensemple. They have used also Fisher linear discriminates (FLD) 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. The local approach can be either based on the geometric properties of the features or some appearance based methods that transform the image with a mathematical representation. Pantic and Rothkrantz (2000) used geometric features to categorize in different

action units as well as combinations of action units and basic emotions. Tian et al. (2001) detected and tracked changes in facial components. The models that they produced included a lip model with 3 states (open, close, tightly closed), an eye model with 2 states (open and closed), brow and cheek models and transient facial features model with 2 states (present or not present). Their categorization is based on action units.

Gabor based transformations are widely used to extract facial appearance changes. It has been shown that simple cells in the primary visual cortex can be modeled by Gabor functions (Daugman 1980, 1985). This solid physiological connection between Gabor functions and human vision has yielded several approaches to feature extraction (Ye et al. 2004) and facial expression recognition (Zhang et al. 1998; Lyons & Akamatsu 1998; Lyons et al. 1999; Gu et al. 2005; Guo & Dyer 2005; Liu & Wang 2006). Moreover, Gabor functions are optimal for measuring local spatial frequencies (Shen & Bai, 2006). Zhang et al. (1998) compared the Gabor function coefficients at the fiducial points location with the coordinates of the fiducial points and concluded that the first represent the face better than the latter. Donato et al. (1999) reported that Gabor functions performed better than any other method used in both analytic and holistic approaches.

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 (Lyons et al. 1999). The way that these fiducial points are identified in an image can either be automatic (Gu et al. 2005) or manual (Zhang et al. 1998; Lyons et al. 1999; Guo & Dyer 2005).

For a more elaborate approach related to facial expression recognition the reader can refer to Pantic and Rothkrantz (2000) and Fasel and Luettin (2003).

Classification

The last step of an AFER system is the classification of the feature vectors into meaningful categories. The distinction between classification methods used in the literature depends on whether or not temporal information is used. Moreover, there is another distinction in terms of the categories that the classifiers classify into being basic emotions, single action units based on FACS or action units combinations that are used to form broader notions of emotions such as fear or stress and so on.

A Hidden Markov Model (HMM) describes the statistical behaviour of a process that generates time series data having certain statistical characteristics. Lien et al. (2000) used the temporal characteristics and HMM to classify into action units or combinations of action units. A comparative study of the performance of different classifiers is provided by Cohen et al. (2003). They have used both static and dynamic classifiers such as Naïve-Bayes based classifiers and Hidden Markov Models (HMM), respectively, to classify into basic emotions. The static classifiers used were, a modified Naïve-Bayes which assumed the distribution to be Cauchy not Gaussian and a Tree-Augmented Naïve Bayes classifier. They have also employed a multi-level HMM which allowed to segment long video sequences to different expression segments using temporal information.

Static classifiers do not use any temporal information that is available in image sequences. They use the information of a single image. Several methods can be found in the literature including neural networks, support vector machines (SVM), etc. Guo and Dyer (2005) provide a comparative study of different classifiers using the simplified Bayes, SVM and combinations of

these classifiers using Adaboost. They also proposed a Linear Programming classifier. They categorized into basic emotions using the JAFFE database. Neural Networks have been deployed in various studies as well known classifiers for multi-class problems (Zhang et al. 1998).

Temporal classifiers are more suitable for person-depended tasks due to their higher degree of variability in expression in humans as well as the variation in the dynamics of each expression. They are considered more difficult to train since they need a larger training set and more parameters in order to train them adequately. Static classifiers can be problematic when they are used in sequences where each frame is categorized. When the expression is not at its peak it is likely that the static classifier can perform poorly. On the other hand static classifiers are easily using a smaller number of parameters

APPLICATION

On this section an approach for automatic facial expression recognition is presented. The proposed methodology includes four stages: (a) automatic discovery of prominent features of a face, such as the eyes, and subsequent 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 2).

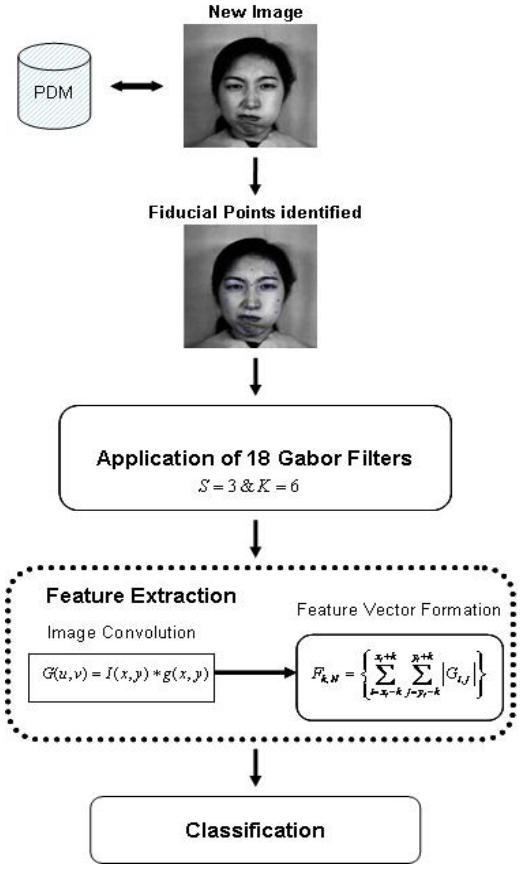

Figure 2 Flow Chart of the proposed method

Active Shape Models

Active Shape Models (Cootes et al. 1995) 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 a single vector is represented 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 Equation (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 Equation (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.

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 which 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.

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 (Dougman 1980,1985; Manjunathan & Ma 1996)

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

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 (Dougman, 1980).

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 and Ma showed that Gabor filters include redundant information in the images produced by the filter (Manjunath & Ma, 1996; Guo & Dyer, 2005). 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:

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

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

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

where a is the scaling factor, S is the number of scales, $m = 0, 1, \ldots, S-1$, U_h and U_l are the high and low frequency of interest. In this work $U_h = \sqrt{2}/4, \quad U_l = \sqrt{2}/16$ are chosen with three scales and six orientations differing by $\pi/6$. A total of 18 different Gabor Filters are defined which are used to extract the feature vector.

Feature Extraction

The Gabor decomposition of any given image at any scale and orientation is produced by convolving the image with a particular filter. The magnitude of the resulting complex image is used to define the features that will form the feature vector. The feature vector is formed according to the following equation:

$$F_{k,l} = \left\{ \sum_{i=x_l-k}^{x_l+k} \sum_{j=y_l-k}^{y_l+k} |G_{i,j}| \right\}, \quad l = 0,1,..,N, \quad k = 0,1,..,5, \tag{10}$$

where N is the number of the fiducial points used, and k is the number of neighboring pixels used to form the regions.

A total of 20 fiducial points are used to form the feature vector and regions of different size are employed to evaluate the methodology.

Artificial Neural Networks

A feed forward back propagation ANN is employed. The architecture of the ANNs is shown schematically in the Figure 3.

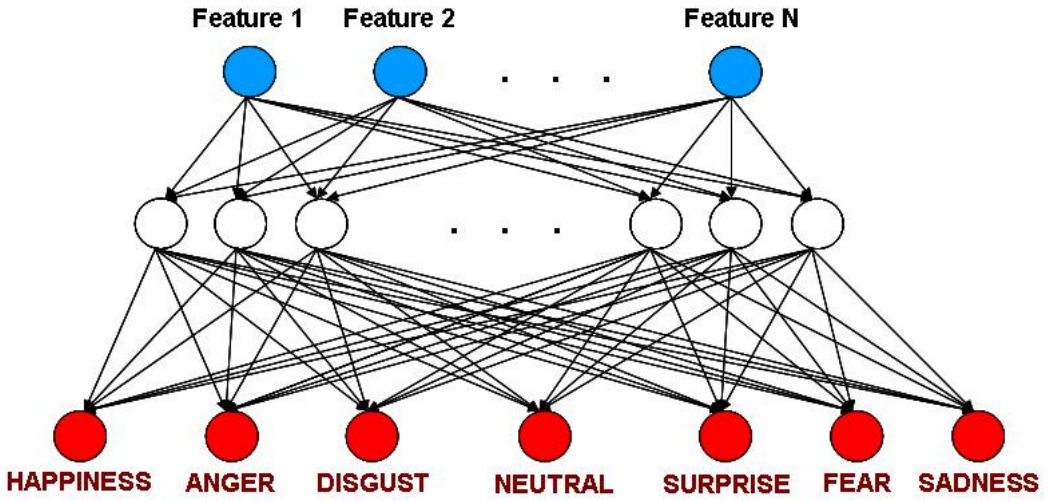

Figure 3 Artificial Neural Network Architecture

The first layer referred to as input layer consists of t inputs which is the dimension of the feature vector. The second layer referred to as hidden layer consists of $t + c/2$ neurons, where c is the number of classes used for classification. Finally, the output layer consists of the basic emotions and the neutral position. The sigmoid function is used as activation function for these hidden neurons. 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.

Dataset

The JAFFE database (Lyons et al. 1999) 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 it is included in the annotation of the database as a seventh basic emotion.

Results

Seven sets of experiments are conducted using automatic identification of fiducial points and are compared with seven sets of experiments conducted when 34 fiducial points are manually identified. Table 1 presents the accuracy of the methodology for both sets of points and all different regions which are used. In the tables presented below 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 that the when

the region gets broader it 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 whereas when 34 points are used the dimension is 612.

Neighborhood 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 1: Accuracy obtained for different region sizes.

The best accuracy is reported when a region of 9x9 pixels is used for the 20 fiducial points set. In Table 2 below 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. (1998) have excluded fear from their experiments due to the difficulty of expressing the emotion from the subjects and some evidence that fear is processed differently by the human brain. Yin et al. (2006) reported difficulties even among human experts to distinguish certain emotional states, namely sad with fear and disgust with anger.

	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

Table 2: Confusion matrix of the best performing region (9x9) for the 20 points set.

Zhang et al. (1998) 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 (2005) 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 proposed by Zhang et al. (1998) which were manually selected. Three classifiers were compared and the accuracy of each are presented. When the Simplified Bayes was used the reported accuracy was 63.3%, when the linear Support Vector Machines (SVM) was used the reported accuracy was 91.4% and when the non linear (Gaussian Radial Basis function kernel) SVM was used the reported accuracy was 92.3%. The methodologies presented above construct the feature vector utilizing information from a single pixel, the pixel that the fiducial point corresponds. This pixel-based approach can be modified to accommodate information from neighboring pixels at the location of each fiducial point forming a neighborhood, named region. The advantage of this modification is twofold: first artefacts that are introduced due to imprecise identification of prominent features of the face are avoided; an automatic methodology is more likely to vaguely identify the exact location of a fiducial point than a human expert. Second, a larger region is utilised which carries more information at certain areas of the face that contain important information on the facial muscle movement, allowing the reduction of the number of the fiducial points used to 20 (14 less that previous approaches). This is a 42% dimensionality reduction at the feature vector allowing for faster computation. The methodology has an accuracy of 90.2% and can be compared with methods that use single-pixel information and more fiducial points that are manually identified.

CONCLUSIONS

Automatic facial expressions recognition is a vital issue in human interpersonal communication. Systems that are able to perform well and analyse facial expressions in real world examples are advantageous for scientific applications as well as everyday real world applications.

In this chapter an approach to automatic facial expression recognition system 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. A region based methodology that ensures some flexibility on the identified points and avoids artefacts is employed. Moreover, a 20 fiducial point set is used that models facial geometry adequately for facial expression recognition. The methodology presented does not perform very well when trying to classify sadness or fear and reports the biggest losses between the two emotions but has been reported in the literature that these emotions often are troubling for human experts also and cannot be adequately distinguished (Zhang et al., 1998;Yin et al., 2006).

FUTURE TRENDS

Automatic facial expression systems will steadily move towards real world applications. In terms of research there are still fields that must be investigated in order to allow the transition of AFER systems to real world applications.

A very persistent requirement is often defined in terms of speed and accuracy of the system. The AFER systems should be developed to operate in real time and to be fully automated without manual intervention. Modern computer systems are close to allow this kind of processing overhead and there are system, usually embedded, that allow to operate in real time. More efficient methods for face identification, recognition and acquisition in terms of speed and accuracy would facilitate the application of AFER systems in real world examples.

An active research field concerning the AFER systems is the categorization of such systems not only in basic, global emotions that are limited in nature, but also in facial actions or deformations that would allow more diversity in terms of the categorized emotions. Basic

emotions cover a small set of the emotions that are present in a human face in every day life. Scientific subjects that would benefit from an active fully working AFER system are very little concerned with basic emotions and study different states and emotions such as pain, stress, fatigue and so on. This will also be beneficial moving towards real-world applications since there is a distinction between posed expressions and spontaneous expressions. The databases that are currently in use in the scientific community do not include data for spontaneous expressions.

ACKNOWLEDGMENT

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

REFERENCES

Black, M.J. & Yacoob, Y. (1998). Recognising facial expressions in image sequences using local parameterised models. Int'l J Computer Vision, 25(1), 23-48.

Cohen, I., Sebe, N., Garg, A., Chen, L.S. & Huang, S.T. (2003). Facial expression recognition from video sequences: temporal and static modelling, Computer Vision and Image Understanding, 91, 160-187.

Cootes, T.F., Edwards, G. & Taylor C.J. (1998). Active appearance models, Proc. European Conf. Computer Vision, 2, 484-498.

Cootes, T.F., Taylor, C.J., Cooper, D.H. & Graham, J. (1995). Active shape models – Their training and application, Computer Vision and Image Understanding, 61(1), 38-59.

Daugman, J. (1980). Two-dimensional spectral analysis of cortical receptive field profiles, Vision Research, 20, 846-856.

Daugman, J. (1985). Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical fields, J Optical Society of America A, 2(7), 1160-1171.

Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P. & Senjowski, T.J. (1999). Classifying facial actions, IEEE Trans. Pattern Analysis and Machine Intelligence, 21(10), 974-989.

Ekman , P. (1991). Telling Lies: Clues to deceit in the Marketplace, Politics, and Marriage. W.W. Norton, New York, USA.

Ekman, P. & Friesen, W.V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124-129

Ekman, P. & Friesen, W.V. (1978). The Facial Action Coding System: A technique for the measurement of facial movement. Consulting Psychologist Press, San Francisco

Ekman, P. & Rosenberg, E.L., (2005). What the face reveals: Basic and applied studies of spontaneous expression using the FACS, Oxford University Press, Oxford, UK.

Ekman, P. (2003). Darwin, deception and facial expression. Annals New York Academy of sciences, 100, 205-221.

Essa, I. & Petland, A. (1997). Coding, analysis, interpretation, recognition of facial expressions, IEEE Trans. Pattern Analysis and Machine Intelligence, 19(7), 757-763.

Fasel, B. & Luettin, J. (2003). Automatic facial expression analysis: a survey, Pattern Recognition, 36, 259-275.

Fasel, I.R., Fortenberry, B. & Movellan, J.R. (2005). A generative framework for real time object detection and classification. Int'l J Computer Vision and Image Understanding, 98(1), 181-210.

Freund, Y. & Schapire, R.E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. Computational Learning Theory: Eurocolt 95, 23–37.

Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. The Annals of Statistics, 28(2), 337-374.

Friesen, W.V. & Ekman, P. (1983). Emfacs-7: emotional facial action coding system, Unpublished Manuscript, University of California at San Francisco

Gu, H., Zhang, Y., & Ji, Q. (2005). Task oriented facial behavior recognition with selective sensing, Computer Vision and Image Understanding, 100, 385-415.

Guo, G. & Dyer, C.R. (2005). Learning From Examples in the Small Sample Case: Face Expression Recognition, IEEE Trans. Sys. Man and Cybernetics–PART B: Cybernetics, 35(3), 477-488.

Hjelmas, E. & Low B.K. (2001). Face detection: A survey, Computer Vision and Image Understanding, 83, 236-274.

Kanade, T., Cohn, J.F. & Tian, Y. (2000). Comprehensive database for facial expression analysis, Proc. IEEE Int'l Conf. Face and Gesture Recognition, 46-53.

Kass, M., Witkin, A. & Terzopoulos, D. (1987). Snakes: Active contours models, Proc Int'l Conference Computer Vision, 259-269.

Li, S.Z. & Jain, A.K. (2005). Handbook of face recognition, Springer, New York, USA.

Lien, J.J., Kanade, T., Cohn, J.F. & Li C.C. (2000). Detection, tracking and classification of action units in facial expression, Robotics and Autonomous Systems, 31, 131-146.

Liu, W. & Wang, Z. (2006). Facial Expression Recognition Based on Fusion of Multiple Gabor Features, Int'l Conf Pattern Recognition,536-539.

Lyons, M. & Akamatsu, (1998). Coding Facial Expressions with Gabor Wavelets, Int'l Conf Automatic Face and Gesture Recognition, 200-205.

Lyons, M.J., Budynek, J., & Akamatsu, S. (1999). Automatic classification of single facial images. IEEE Trans. Pattern Analysis and Machine Intelligence, 21(12), 1357-1352.

Manjunath, B.S. & Ma, W.Y. (1996). Texture features for browsing and retrieval of image data, IEEE Trans. Pattern Analysis and Machine Intelligence, 18(8), 837-842.

Meihle, A. (1973). Surgery of the facial nerve. Saunders, Philadelphia, USA.

Pantic, M. & Rothkrantz L. (2000). Automatic analysis of facial expressions: The state of the art, IEEE Trans. Pattern Analysis and Machine Intelligence, 22(12), 1424-1445.

Pantic, M. & Rothkrantz, L. (2000). Expert system for automatic analysis of facial expressions, Image and Vision Computing, 18(11), 881-905.

Pantic, M., Valstar, M.F., Rademaker, R. & Maat. L. (2005). Web-based database for facial expression analysis, Proc. IEEE Int'l Conf. Multimedia and Expo, 317-321.

Petland, A., Moghaddam, B. & Starner, T. (1994). View-based and modular eigenspaces for face recognition, Proc IEEE Conf. Computer Vision and Pattern Recognition, 84-91.

Rowley, H., Baluja, S. & Kanade, T. (1998). Neural network-based face detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 20(1), 23-38.

Schneiderman, H. & Kanade, T. (2000). A statistical model for 3d object detection applied to faces and cars, Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 746-751.

Shen, L. & Bai, L. (2006). A review on Gabor wavelets for face recognition, Pattern Analsysis and Applications, 9, 273-292.

Sung, K.K. & Pogio, T. (1998). Example-based learning for view-based human face detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 20(1), 39-51.

Sunko, F.M., Ordaas, S., Butakoff, C., Cruz, S. & Frangi, A.F. (2007). Active shape models with invariant optima features: Application to facial analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, 29(7), 1105-1117.

Tian, Y., Kanede, T. & Cohn J.F. (2001). Recognizing action units for facial expression analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, 23(2), 97-115.

Tian, Y.L., Kanade, T. & Cohn J.F. (2001). Recognizing action units for facial expression analysis. IEEE Trans. Pattern Analysis and Machine Intelligence, 23(2), 97-115.

Van Ginneken, B., Frangi, A.F., Staal, J.J., Ter Har Romeny, B.M. & Viergever, M.A. (2002). Active shape model segmentation with optimal features, IEEE Trans. Medical Imaging, 21(8), 924-933.

Viola, P. & Jones, M. (2001). Rosubt real-time object detection, Int'l Workshop on Statistical and Computational theories of Vision - Modeling, Learning, Computing and Sampling.

Viola, P. & Jones, M. (2004). Robust real-time face detection. J. Computer Vision, 57(2), 137-154.

Yang, M.H., Kriegman, D.J. & Ahuja, N. (2002). Detecting faces in images: a survey, IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(1), 34-58.

Ye, Z., Zhan, Y. & Song, S. (2004). Facial expression features extraction based on Gabor wavelet transformation, IEEE Int'l Conf Systems, Man and Cybernetics, 2215-2219.

Yin, L., Wei, X., Sun, Y., Wang, J. & Rosato, M. (2006). A 3D facial expression database for facial behavior research, Proc. IEEE Int'l Conf. Face and Gesture Recognition, 211-216.

Zhang, Z., Lyons, M., Schuster, M. & Akamatsu, S. (1998). Comparison Between Geometry-Based and Gabor-Wavelet-Based Facial Expression Recognition Using Multi-Layer Perceptron, Int'l Conf Automatic Face and Gesture Recognition, 454-459.

KEY TERMS & DEFINITIONS

Action Unit (AU) – Is the key element of FACS, each action unit describes facial deformation due to each facial muscle movement. There are a total of 44 AUs where the majority involves contraction or relaxation of facial muscles and the rest involve miscellaneous actions such as “tongue show” or “bite lip”.

Basic Emotions – They are a small set of prototypic emotions which share characteristics of universality and uniformity across people with different ethnic background or cultural heritage. The six basic emotions were proposed by Ekman and Friesen (1971) and are: disgust, fear, joy, surprise, sadness and anger.

Classification – Is the task that categorizes feature vectors into appropriate categories. Each category is called a class.

Facial Action Coding System (FACS) – It is a system developed by Ekman and Friesen (1978) to categorize human expressions. Using FACS human coders can categorize all possible facial deformation into action units that describe facial muscle movement.

Feature vector extraction – Is the task of providing a feature vector that describes facial geometry and deformation. There are two ways to model facial geometry and deformation: first by using prominent features of the face and second by using a mathematical transformation so that changes in appearance are modeled.

Image Processing – The analysis of an image using techniques that can identify shades, colors and relationships which cannot be perceived by the human eye.

Machine Learning –The purpose of machine learning is to extract information from several types of data automatically, using computational and statistical methods. It is the use of computer algorithms which improve automatically using experience

Point Distribution Model (PDM) – It is a model that tries to form a distribution of sample points from the training set. When the PDM is constructed it can approximate the position of each model point in a new image without manual intervention.