Neural Network Classification of Photogenic Facial Expressions Based on Fiducial Points and Gabor Features

Luciana R. Veloso, João M. de Carvalho, Claudio S.V.C. Cavalvanti, Eduardo S. Moura, Felipe L. Coutinho, and Herman M. Gomes

Universidade Federal de Campina Grande,
Av. Aprigio Veloso s/n, 58109-970 Campina Grande PB
{veloso,carvalho}@dee.ufcg.edu.br
{csvcc,edumoura,felipelc,hmg}@dsc.ufcg.edu.br

Abstract. This work reports a study about the use of Gabor coefficients and coordinates of fiducial (landmark) points to represent facial features and allow the discrimination between photogenic and non-photogenic facial images, using neural networks. Experiments have been performed using 416 images from the Cohn-Kanade AU-Coded Facial Expression Database [1]. In order to extract fiducial points and classify the expressions, a manual processing was performed. The facial expression classifications were obtained with the help of the Action Unit information available in the image database. Various combinations of features were tested and evaluated. The best results were obtained with a weighted sum of a neural network classifier using Gabor coefficients and another using only the fiducial points. These indicated that fiducial points are a very promising feature for the classification performed.

Keywords: facial expression analysis, Gabor coefficients, facial fiducial points, neural networks.

1 Introduction

A major task for the Human Computer Interaction (HCI) community is to equip the computer with the ability to recognize the user's affective states, intentions and needs from a set of non-verbal cues, hence significantly enhancing the interaction between human and machine.

One of the most difficult investigated tasks in this area is the recognition of the emotional state of its users. Many systems have been investigated and proposed, both by industry and by the scientific community, with this objective. This effort is motivated by the relevance that emotional expression have for human communication [2], constituting the most powerful, natural, and immediate means for human beings to communicate their emotions and intentions. Mehrabian affirms in this article [3] that facial expressions contributes for 55% to the effect of the message as a whole, while voice intonation and verbal contributes with 38% and 7%, respectively.

Despite the large number of possible facial expressions, only a small set of prototype emotional expressions have been investigated: joy, sadness, surprise,

D. Mery and L. Rueda (Eds.): PSIVT 2007, LNCS 4872, pp. 166–179, 2007.

anger, fear and disgust. This categorization was first discussed by Darwin in 1872 [4] and is generally accepted by psychologists working on facial expression analysis. Although there is some objection to the idea that these expressions signal similar emotions in people of different cultures, it is conceded the cross-cultural consistency of the combinations of facial movements (behavioral phenotypes) that compose the six basic expressions.

In this work, we investigate the relationship between emotional face expressions and the concept of a photogenic expression which is new to the Computer Vision literature, but well known to the Photography field. The concept is normally associated with the attractiveness of a person as a subject for photography. A neural network classifier was utilized to discriminate from photogenic (labeled from neutral and happy emotions) and non-photogenic (labeled from disgust, anger, sadness, fear and suprise emotions) facial expressions.

Within the above context, this work reports a study about the viability of using a set of the fiducial points extracted from a face image and represented by their (x, y) coordinates and gabor features to classify photogenic and non-photogenic facial expressions. Section 2 presents a review of related work and Section 3 describes the methodology employed in the work. Section 4 describes the experiments performed and results obtained. Conclusions and proposals of further research are presented in Section 5.

2 Related Work

The photogeny classification problem has been firstly addressed in a paper [5] that has been chosen for discussion in the next paragraph. Nonetheless, there is a number of other related work on general facial expression recognition that is equally worth discussing in this section.

Batista et. al [5] explored an appearance-based approach to photogenic expression discrimination. PCA features extracted from the grey level information of the face images has been used as input to a SVM and a MLP neural network classifiers. The best recognition rate for an experiment involving re-labelled images from the Cohn-Kanade database [1], when using a MLP neural network classifier, was 87.5%. In the present paper, a completelly different set of features (fiducial points) has been investigated with promissing results. The next paragraphs will discuss the more general facial recognition related work.

Essa and Pentland [6] presented the results of facial expressions recognition based on optical flow, coupled with geometric, physical and motion-based face models. They used 2D motion energy and history templates that encode both the magnitude and the direction of motion. By learning the ideal 2D motion views for four emotional expressions (anger, disgust, happiness and surprise), they defined spatio-temporal templates for those expressions, from which recognition can be performed by template matching. Although the approach proposed by Essa and Pentland is not fully validated, it constitutes an unique method for facial expressions classification.

Kanade et al. [7] proposed a system that recognizes Action Units (AUs) and AUs combinations [8] in facial image sequences using Hidden Markov Models.

Initially, a manual marking of facial feature points around the contours of the eyebrows, eyes, nose and mouth in the first frame of an image sequence is performed. Next, the LucasKanade optical flow algorithm is used to track automatically the feature points in the remaining frames. In the case of the upper face, the WuKanade dense optical flow algorithm and high gradient component detection are used to include detailed information from the larger region of the forehead.

Texture appearance provides an important visual cue to classify a variety of facial expressions. Working on this hypothesis Wang and Yin [8] presented an approach to represent and classify facial expressions based in Topographic Context (TC). Topographic Context describes the distribution of topographic labels in the face region of interest. The face image is split into a number of expressive regions and the facial topographic surface is labeled to form a terrain map. Statistics of the terrain map are then extracted to derive the TC for each of the pre-defined face expressive regions. Finally, a topographic feature vector is created by concatenating the TCs of all expressive regions. With the extracted TC features, facial expressions were recognized using four classifiers: quadratic discriminant classifier - QDC, linear discriminant analysis - LDA, nave Bayesian network classifier - NBC and support vector classifier SVC. The LDA classifier achieved the best average result of 82.68% correct recognition rate.

Lanitis el al. [9] proposed an unified approach to problems of face image coding and interpretation, using flexible models which represent both shape and gray-level appearance. For shape models, the shapes of facial features and the spatial relations between them are captured in single models, using 152 coordinates of landmarks points in the face. For the training examples, those points were manually located. The model can accurately approximate the shape of any face in the training set using 16 shape parameters (eigenvectors weights). The shape-free gray level appearance model was obtained by deforming each face image to match a mean shape in such a way that changes in gray-level intensities are kept to a minimum. Landmarks (14 in total) were used to deform the face images and gray-level intensities within the face area were extracted. Only 12 variables were needed to account for 95% of the training set variation in the flexible gray-level model. The last model, local gray-level appearance, was built from the projection profile at each model point. The shape model a the local gray-level models can be used to automatically locate all the modeled features, using Active Shape Models search. The facial expression recognition problem was investigated by establishing the distribution of appearance parameters over a selected training set for each expression category so that the appearance parameters calculated for a new face image could be used for determining the expression. The best results were obtained with shape-free gray level. The recognition rate was 74%, with shape-free gray level, 70%, with shape + shappe-free model and 53%, with shape model.

In the work of Zhang et. al[10], geometric positions of a set of fiducial points and Gabor wavelet coefficients were applied to recognize facial expressions. Each image was convolved with 18 Gabor filters, comprising 3 scales and 6 orientations at the location of the fiducial points. Therefore, the images were represented

by a vector of 612 (18X34) elements, each. The classifier architecture is based on a two-layer perceptron. Experiments were performed using 10-fold cross-validation. Gabor filters reached a better recognition rate than that obtained using only geometric positions (coordinates of the points). The recognition rate was 73.3%, with geometric positions, 92.2%, with Gabor filters and 93.3%, with combined information.

An expert system for automatic analysis of facial expressions, called Integrated System for Facial Expression Recognition (ISFER), was developed by Pantic and Rothkrantz [11]. The system consists of two parts. The first part utilizes a parallel of multiple techniques: detention of nose and chin, curve fitting of the eyebrow, chain code eyebrow, neural network approach to eye tracking, adjustment of curves for mouth detention, detector fuzzy of mouth and search of profile contour. The second part of the system is its inference engine called HERCULES, which converts low level face geometry into high level facial actions, and then this into highest level weighted emotion labels.

For model building, shapes are usually represented by a set of reference points, sometimes called fiducial points, taken from well defined image edges. The most direct way of obtaining these points is by hand marking, on all images of the training and test sets. Although quite simple, this procedure is a rather tedious and error prone task. Besides modeling, reference points are also used for face and facial expressions recognition tasks, as they allow detection of significant face features. Much work has been announced on the development of systems for automatic marking, or annotation, of reference points, but no efficient method for that is thus far available.

Locating and tracking facial features in image sequences is the objective of the work presented by Zhu and Ji [12]. An Adaboost classifier is initially utilized for face and eye detection in an initial image. Eyes location is then used to position a trained face mask, which allows a first (rough) detection of 28 fiducial points. Next, a vector of Gabor coefficients is calculated at each detected point and compared (similarity search) to vectors in the training set. The most similar vectors in the training set provide new estimates for the fiducial points, and so on, until convergence is achieved. Appearance parameters and geometric relationships from the extracted facial features are utilized by a mechanism based on correlation and constraints of face shape to track the fiducial points in subsequent image frames.

A probabilistic approach based on multi-modal models has been proposed by Tong and Ji [13], aiming to capture relationships between features extracted from different face view angles, using PCA. An eye detector is initially utilized to determine starting points for a feature point search procedure. A set of multi-scale and multi-directional wavelets is employed for local appearance modeling around the detected feature points. An EM algorithm is finally utilized to estimate multi-modal parameters. Experimental results show that this technique performs well for facial features tracking in cases of large pose variations.

In this paper, the geometric positions of a set of fiducial points and Gabor wavelet coefficients, which was initially utilized by Zhang [10], has been investigated

to the photogenic classification problem. Other differeces of the proposed approach when compared to previous work is the small number of fiducial points utilized and the combination of classifiers - one using Gabor coefficients and the other using the fiducial points themselves.

3 System Description

Our main goal was to train a classifier to learn the relationships between emotional face expressions and the concept of a photogenic expression, normally associated with a good picture of a person.

In this section a system is proposed to perform facial expression recognition from facial fiducial points, using a Neural Network (NN) classifier. The system's architecture is composed of four main modules: fiducial point extractor, Gabor wavelet extractor, mask normalization and Neural Network classifier (NN). The fiducial point extraction module is responsible for locating and extracting the coordinates of discriminant (landmark) points on each face. In the next module, the Gabor wavelet extractor, a set of Gabor coefficients are extracted at each fiducial point. The fiducial points are normalized in terms of Cartesian origin and face orientation by the mask normalization module. Each image is represented by two feature sets: the coordinates of the fiducial points and the coefficients of the Gabor wavelets. After normalization, the two sets are used to train the NN module which will perform the recognition of facial expression patterns.

3.1 Fiducial Points

For the present work, the fiducial points were hand marked for all images in the Conh-Kanade database [1]. The development of an automatic fiducial points extractor is currently under way, using the Active Appearance Model (AAM) technique (see Appendix 5) . All face images in the database have been previously and originally labelled as belonging to one of the following facial expressions: happiness, sadness, anger, fear, surprise, disgust and neutral. For our work, these labels have been mapped into just two labels: photogenic and non-photogenic. On each face, 29 points are marked, as illustrated in Figure 1.

A software tool was developed with the objective of facilitating the annotation of fiducial points and class labeling for the images in the database. This tool was named Face Descriptor. The Face Descriptor incorporates a face detector and a graphical interface. Upon examination of its facial features, each face in a given image is labelled as belonging to following categories: happiness, anger, fear, surprise, sadness, disgust and neutral.

With the Face Descriptor, the user can either work with a previously annotated and labeled image or perform those operations on a new image. For new images, the menu option "File Open image without XML must be selected, which will prompt the face detected (if there is one) in the opened image to be shown within a square, as illustrated in Figure 2. Annotation (markings) of the fiducial points is done in a pre-defined order, by just positioning the cursor on

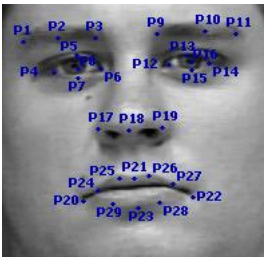

Fig. 1. Image with marked fiducial points

the selected point and pressing the mouse button, which causes the point coordinates to be stored in the propper field in the "Facial Regions Coordinates frame (Figure 2). As an aid to the user, a reference face is displayed in the interface, indicating the next point to be marked.

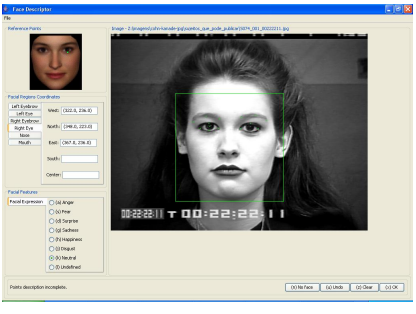

Fig. 2. Face Descriptor interface window

Face labeling is achieved by selecting the appropriate menu options in the "Facial Features frame, also shown in Figure 2. Once the labeling is finished, the "ADD FACE button is pressed to indicate that the processing of the current face is concluded. If other faces are present in the analyzed image they will be automatically detected and displayed for further marking and labeling. When a not valid face is detected, the "NO FACE button has to be pressed, to discard that detection and proceed to the next. When no more faces are detected, the user presses the "OK button to conclude the operation, which prompts the "Face Descriptor software to save the selected points coordinates and labels in XML (eXtensible Markup Language) files, one file for each face. The "UNDO button is used to discard wrongly marked points. The "CLEAR button erases all saved information, allowing the user to start all over again. "The File Exit menu option closes the software.

For each marked point, cartesian coordinates are extracted and stored in a XML file. Table 1 shows the image quantities used for each facial exression. The set of the extracted points forms a mask, representing the face image.

Table 1. Image distribution among expressions

Expression	Number of Images
Anger	92
Fear	128
Surprise	156
Sadness	150
Hapiness	205
Disgust	84
Neutral	537

3.2 Gabor Wavelets

Gabor wavelets are receiving great attention in the area of Facial Expression Recognition [14] [15] [10]. The Gabor wavelets capture the properties of spatial localization, orientation selectivity, spatial frequency selectivity and quadrature phase relationship [16]. In the face image, the convolution of the image with a family of Gabor filters produces salient local features, such as eyes, nose and mouth. The two-dimensional Gabor function describes a sinusoid of frequency W modulated by a Gaussian:

$$g(x,y) = (\frac{1}{2\pi\sigma_x\sigma_y})e^{[\frac{-1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})+2\pi(-1)^{1/2}W_x]} \quad (1)$$

where σ_x and σ_y are the widths of the Gaussian in the spatial domain, that is, along x and y axis.

The Gabor functions form a complete, but non-orthogonal, basis set and any given function $f(x,y)$ can be decomposed in terms of these basis functions. Such a decomposition results in a family of Gabor filters, making possible to detect features at various scales and orientations.

In this work, a set of multi-scale and multi-orientation Gabor Wavelets are employed to model local appearances around ten fiducial points, localized around the eyes and the mouth, except the eyes inner corners. Each image is convolved with four orientation filter and two frequencies filters. Therefore, the vector of the characteristic is composed of 80 Gabor wavelet coefficient, 8 coefficients at each fiducial point analyzed.

3.3 Mask Normalization

In the third module of the system, all fiducial point coordinates are normalized by the length of the line segment joining the fiducial points corresponding to the eyes inner corners (this length becomes unitary). Additionally, mask orientation is

normalized (slope correction) and the origin of the Cartesian coordinates system is translated to the middle point of that line segment, i.e., to the middle point between the eyes. These operations are depicted in Equations 2-5.

$$x' = x - x_c \tag{2}$$
$$y' = y - y_c \tag{3}$$
$$x' = x * \cos\theta - y * \sin\theta \tag{4}$$
$$y' = y * \cos\theta + x * \sin\theta \tag{5}$$

where (x', y') and (x, y) are the normalized and original coordinates, respectively, (x_c, y_c) are the coordinates of the middle point between the eyes, and θ is the slope (angle) of the line joining the eyes inner corners, used for slope correction.

3.4 Neural Network Training

All neural network trainings were performed using MLP (Multilayer Perceptron), with varying sizes of input and hidden layers. The sizes utilized in the experiments have been experimentally discovered as a function of the type and amount of input features. In total, five types of neural networks were used: two neural networks receiving only point coordinates, one neural network receiving Gabor features, another for Gabor features and point coordinates and a last one receiving a set of Gabor features, point coordinates and PCA (Principal Component Analysis) features.

4 Experiments and Results

The experiments with photogenic versus non-photogenic faces were performed using 416 images from the Cohn-Kanade AU-Coded Facial Expression Database [1]. From the total image set, 208 images were labeled as photogenic and 208 as non-photogenic. The subset was separated in training 50%, validating 25% and testing 25% further. Figure 3 shows some exemples of this image set.

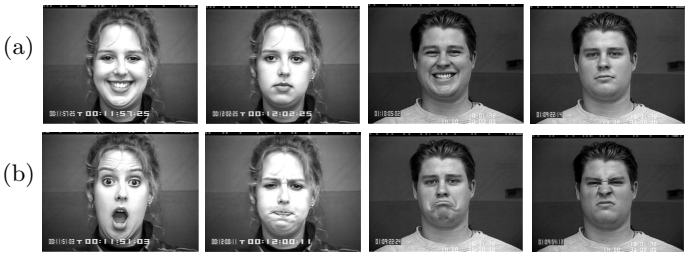

Fig. 3. Examples of (a) Photogenic and (b) Non-photogenic faces

Facial expression classification experiments were performed using the following five types of feature sets: 1)coordinates of all 29 fiducial points (58 coordinates); 2) only eye and mouth fiducial points coordinates (32 coordinates); 3) 10 Gabor features calculated for eight fiducial points of each face (80 features); 4) Gabor features combined with all fiducial points coordinates (138 features); 5)78 features obtained by applying Principal Component Analysis (PCA)[17] to the previous feature set (4). The main goal of PCA is to reduce the dimension of a data set while retaining as much information as possible. Essentially, a set of correlated variables is transformed into a set of uncorrelated variables (by linear combination) and ordered by decreasing variability. The resulting set of components in this work account for 90% of the total data variance.

A different neural network (NN) was trained for each type of feature set/ experiment. Due to the fact that random weight initialization can lead to different error energy minima, each experiment (training and testing) was repeated 10 times, using a Intel Xeon 5130 Processor - Dual core / 2.00 GHz / 1333MHz FSB with 4GB of RAM. Training time varied between 843,3 and 14,5 seconds, depending on the features set size and initial parameters.

Table 2 summarizes the experiments results. It shows the number of nodes at the three (input, hidden and output) layers of the NN, as well as the maximum, mean and standard deviations values obtained for the correct classification rates, claculated from the 10 repetitions of each experiment type. The average classification (execution) time is also shown in Table 2.

Table 2. Results for the five types of experiments performed. Input, Hidden and Output are the number of nodes at each layer of the Neural Network. Max, Mean and STD are the Maximum, Mean, and Standard Deviation for the classification rates, respectively, calculated for 10 repetitions of each experiment. Execution time (in seconds) is the average time (for 10 repetitions) required by each NN to produce a classification result.

Type	Input	Hidden	Output	Max	Mean	STD	Exec. time
1	58	29	2	75.50	74.60	0.65	$9.2 * 10^{-4}$
2	32	16	2	73.50	71.70	0.90	$6.6 * 10^{-4}$
3	80	40	2	68.50	64.60	2.21	$7.2 * 10^{-4}$
4	138	69	2	72.00	69.00	1.29	$1.2 * 10^{-1}$
5	78	38	2	71.00	68.70	1.28	$1.4 * 10^{-1}$

4.1 Classifiers Fusion

The idea of combining classifiers in order to compensate for their individual weakness and to enhance their individual strengths has been widely used in recent pattern recognition applications [18]. Using this approach, different types of features can be independently used to classify a given pattern and the classifiers outputs combined to achieve an overall performance which is better than that of any individual classification.

This section presents a multiple classifiers algorithm based on two different photogenic classifiers: Coordinates-NN (type 1 on Table 2) and Gabor-NN (type 3 on Table 2). For this hybrid classifier we needed to define a combination rule for the classifiers outputs. In this work, three combining strategies have been considered. Initially, assume that an object Z must be assigned to one of the possible classes and that a number of classifiers are available, each representing the given pattern by a distinct measurement vector. Denote the measurement vector used by the *ith* classifier as x_i and the a posteriori probability $P(w_j|x_i, ..., x_l)$. Therefore, the combination rules are:

Sum (S): Assigns Z to class if

$$\Sigma_{i=1}^{L} P(w_j|x_i) = max_{k=1}^{K} \Sigma_{i=1}^{L} p(w_k|x_i) \tag{6}$$

Product (P): Assigns Z to class if

$$\Pi_{i=1}^{L} P(w_j|x_i) = max_{k=1}^{K} \Pi_{i=1}^{L} P(w_k|x_i) \tag{7}$$

Weighted sum (WS): Assigns Z to class if

$$\Sigma_{i=1}^{L} \alpha_i P(w_j|x_i) = max_{k=1}^{K} \Sigma_{i=1}^{L} \alpha_i p(w_k|x_i) \tag{8}$$

where , are weights for the classifiers.

To guarantee that the classifier outputs represent probabilities, output normalization is performed:

$$P^*(w_j|x_i) = \frac{P(w_j|x_i)}{\Sigma_k P(w_j|x_i)} \tag{9}$$

For the weighted sum rule, for which the optimum weights were obtained by an exhaustive search procedure where for each classifiers combination, 2,000 different weight vectors with random adaptation are tested. The average recognition rates obtained considering the two different classifiers combination (types 1 and 3) are as follows: 77.30%, 77.50% and 78.00%, for the sum, product and weighted sum rules, respectivelly. Thus, the best classification rate was achiecved when using the weighted sum rule. Table 3 shows the confusion matrix of the best results of the experiments combining the classifiers. Note that the false acceptance and false

Table 3. Confusion matrix of the best results for the experiments combining the classifiers

	Photogenic	Non-photogenic
Photogenic	77	21
Non-photogenic	23	79

Table 4. Results for Gabor-NN and Coordinates-NN phtogenic classifier with a lower learning rate. The classification rates are expressed in %.

Classifier	Result
Gabor-NN	68.50
Coord-NN	81.00
Combination	82.00

rejection errors are similar (23 and 21). A new training was performed for Gabor-NN and Coordinates-NN photogenic classifiers with a lower learning rate. This was done aiming a better recognition rate to the system (see Table 4).

5 Conclusion

Preliminary experimental results indicate the potencial of fiducial points to distinguish between photogenic/non-photogenic facial expressions. A combination of classifiers performed slightly better (just 1%) than the best classifier investigated (neural network using a set of fiducial points). This can be attributed to the poor performance of the other combined classifier - Gabor-NN (with only 68.50% of correct recognition). Further investigation on this issue will include the investigation of techniques for automatic fiducial point extraction. In a preliminary study towards that direction, we compared the points identified by an AAM (Active Appearance Model) with the available manually marked fiducial points (see the Appendix at the end of the paper). The results showed that, for the most significant points for facial expression analysis, the compared coordinates were not significantly distinct from each other. This is a good indication that the approach will not suffer degradation when using automatically located points. The next step is exactly to redo all performed experiments, this time with the fiducial points automatically located.

Acknowledgments. This work was developed in collaboration with HP Brazil R&D. The authors would like to thank Professor Jeffrey Cohn for granting access to Cohn-Kanade AU-coded Facial Expression Database.

References

1. Cohn, J.F., Zlochower, A., Lien, J., Kanade, T.: Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. Psychophysiology 36, 35–43 (1999)
2. Pantic, M., Rothkrantz, L.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1424–1445 (2000)
3. Mehrabian, A.: Communication without words. Psychology Today 2(4), 53–56 (1968)
4. Darwin, C.: The Expression of the Emotions in Man and Animals. Appleton and Company, New York (1872)

5. Batista, L.B., Gomes, H.M., Carvalho, J.M.: Photogenic facial expression discrimination. In: International Conference on Computer Vision Theory and Applications, pp. 166–171 (2006)
6. Essa, I.A., Pentland, A.P.: Coding analysis interpretation and recognition of facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 757–763 (1997)
7. Cohn, J., Zlochower, A., Lien, J., Kanade, T.: Feature-point tracking by optical flow discriminates subtle differences in facial expression. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 396–401 (1998)
8. Wang, J., Yin, L.: Static topographic modeling for facial expression recognition and analysis. Computer Vision and Image Understanding 108(1-2), 19–34 (2007)
9. Lanitis, A., Taylor, C.J., Cootes, T.F.: Automatic interpretation and coding of face images using flexible models. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 743–756 (1997)
10. Zhang, Z., Lyons, M., Schuster, M., Akamatsu, S.: Comparison between geometry-based and gabor wavelets based facial expression recognition using multi-layer perceptron. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 454–461 (1998)
11. Pantic, M., Rothkrantz, L.J.M.: Expert system for automatic analysis of facial expression. Image and Vision Computing 18, 881–905 (2000)
12. Zhu, Z., Ji, Q.: Robust pose invariant facial feture detection and tracking in real-time. In: International Conference on Pattern Recognition, vol. 1, pp. 1092–1095 (2006)
13. Tong, Y., Ji, Q.: Multiview facial feature tracking with a milti-modal probabilistc model. In: International Conference on Pattern Recognition, vol. 1, pp. 307–310 (2006)
14. Bartlett, M., Littlewort, G., Braathen, B., Sejnowski, T., Movellan, J.: A prototype for automatic recognition of spontaneous facial actions. Advances in Neural Information Processing Systems 15, 1271–1278 (2002)
15. Tian, Y.: Evaluation of face resolution for expression analysis. In: Computer Vision and Pattern Recognition Workshop, pp. 82–82 (2004)
16. Lin, D.T., Yang, C.M.: Real-time eye detection using face-circle fitting and dark-pixel filtering. In: IEEE International Conference on Multimedia and Expo, vol. 2, pp. 1167–1170 (2004)
17. Haykin, S.: Neural Networks: A comprehensive Foundation, 2nd edn. Prentice-Hall, Englewood Cliffs (1998)
18. Carvalho, J.M., Oliveira, J., Freitas, C.O.A., Sabourin, R.: Handwritten month word recognition using multiple classifiers. In: Brazilian Symposium on Computer Graphics and Image Processing, pp. 82–89 (2004)
19. Cootes, T.F., Cooper, D.H., Graham, J.: Active shape models- their training and application. Computer Vision and Image Understanding 61(1), 38–59 (1995)
20. Cootes, T.F., Taylor, C.J.: Statistical models of appearance for computer vision. Technical report, University of Manchester, UK, Imaging Science and Biomedical Engineering (2004)

Appendix: Comparing Manual and Automatic Fiducial Point Extraction

The Active Appearance Model (AAM) approach, developed by Cootes et al. [19] [20], was used for automatically locating fiducial points in untrained faces of the available dataset. After generating the Active Appearance Model from a training face set, the AAM software can be used to automatically detect the

fiducial points. This is performed by a search process, which searches within the image for points that best matches with the learned model. This process, however, is done through a graphical user interface which searches for points on each image at once. In order to speed up experiments, the modeling software was used to provide the fiducial points only for the test face images. Facial images under several expressions have been used for training the AAM with the purposed of obtaining a face model. From this model, the software can extract, for each image, the fiducial points, as described before. Three different metrics have been used to compare the points: the mean of the Euclidean distances between automatic and manual points for (a) individual points; (b) all points; and (c) the Root Mean-Squared Error (RMSE).

In the first evaluation, the mean and the standard deviation of Euclidean distances for each point were computed according to the equation below:

$$\forall fid \mid 1 \leq fid \leq 29 \tag{10}$$

$$mean_{fid} = \sum_{i=1}^{nImages} \frac{d(m_{i,fid}, c_{i,fid})}{nImages} \tag{11}$$

where m is the set of manually selected fiducial points, c is the set of automatically selected fiducial points, $nImages$ is a number of images, fid is fiducial points and d(a,b) is

$$d(a,b) = \sqrt{(a.x - b.x)^2 + (a.y - b.y)^2} \tag{12}$$

In the second evaluation, the mean and the standard deviation of all Euclidean distances were computed as follows:

$$sumAllPoints = \sum_{i=1}^{nImages} \sum_{fid=1}^{29} \frac{d(m_{i,fid}, c_{i,fid})}{nImages} \tag{13}$$

$$meanPoints = \frac{sumAllPoints}{nImages \cdot fid} \tag{14}$$

The standard deviation was evaluated similarly to the mean, but due to space constraints it is not shown here.

In the third evaluation, the Root Mean-Squared error was calculates according to:

$$\forall fid \mid 1 \leq fid \leq 29 \tag{15}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (m_{i,fid} - c_{i,fid})^2}{n}} \tag{16}$$

where n is the number of images. The RMSE can be computed individually for each fiducial point or for all points as in the Equation 17.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \sum_{fid=1}^{29} (m_{i,fid} - c_{i,fid})^2}{n}} \tag{17}$$

The results of the evaluation of the mean and standard deviation of Euclidean distances are shown on Table 5. The experiment was run taking 45 unseen images (5% of the total used) and detecting 29 fiducial points in each image. The distances shown are measured in pixels and the images have the dimensions of 640x490 pixels. A value of 7 pixels in distance has been used to decide that the automatic detection was too divergent from the manual one in a given point. Within this scenario, only the points P1, P9, P10 and P11 were considered too divergent. This can be easily explained since all of them are from eyebrows: P1 is from the left eyebrow and the remaining ones are from the right eyebrow. Eyebrows are slightly more difficult to manually locate than other facial points due to variations in width and texture within the image dataset. These variations may degrade the quality of the trained AAM model. Despite this fact, in an automatic fiducial point extraction scenario, the proposed approach would not be invalidated for two reasons: first, the eyebrows locations are not the main information used for facial expression recognition, and, second, the detected differences are not very significant given the image dimensions. The mean and the standard deviation of all Euclidean distances were 4.88 and 4.20 pixels receptively. The Root Mean-Squared error was 4.55.

Table 5. Mean and the standard deviation of the euclidean distances for each individuals point

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}
Mean	6.98	5.83	6.14	5.07	3.48	4.42	3.95	3.35	7.44	7.79	8.16	3.66	2.70	4.48	3.56
StdDev	4.40	3.21	4.99	4.64	2.53	2.61	2.92	2.64	6.55	7.50	5.50	2.57	1.95	4.43	2.34
	P_{16}	P_{17}	P_{18}	P_{19}	P_{20}	P_{21}	P_{22}	P_{23}	P_{24}	P_{25}	P_{26}	P_{27}	P_{28}	P_{29}	
Mean	2.76	6.43	4.54	5.37	4.61	2.57	4.35	6.54	4.83	3.37	3.37	3.41	5.89	6.39	
StdDev	2.08	5.55	3.10	3.93	2.56	1.43	2.40	6.10	2.98	2.11	2.56	3.23	4.48	4.13	