14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

Enhancing facial expression classification by information fusion

Ioan Buciu, Nikos Nikolaidis and Ioannis Pitas
Department of Informatics
Aristotle University of Thessaloniki
GR-54124, Thessaloniki, Box 451, Greece
Email: {nelu,nikolaid,pitas}@aiia.csd.auth.gr

Alice Caplier and Zakia Hammal
Laboratoire des Images et des Signaux
Institut National Polytechnique de Grenoble
38031 Grenoble, France
Email: alice.caplier@inpg.fr

***Abstract*— The paper presents a system that makes use of the fusion information paradigm to integrate two different sorts of information in order to improve the facial expression classification accuracy over a single feature based classification one. The Discriminant Non-negative Matrix Factorization (DNMF) approach is used to extract a first set of features and an automatically geometrical-based feature extraction algorithm is used for retrieving the second set of features. These features are then concatenated into a single feature vector at feature level. Experiments showed that, when these mixed features are used for classification, the classification accuracy is improved compared with the case when only one type of these features is used.**

I. Introduction

Studied for decades by psychologists for its important role played inside the community, nowadays, facial expression issue knows an increasing interest from the computer scientists community. From the psychology perspective, an emotion expressed by facial features deformation contributes to the communication between humans and can help in cases when the verbal communication is not sufficient or impossible to be performed. Such case appears when, for example, lip reading allows to improve the understanding of a noisy vocal message and it is also a support of communication with hard of hearing people. As far as the computer scientists are concerned, their efforts are focusing toward creating a more friendly human-computer interface which is able to recognize human facial expression and act accordingly. In principle, facial expression classification methods can be divided into three categories: statistical methods [1] that use characteristic points or characteristic blocks in the face; template based methods [2] using models of facial features or models of facial motion and rule based methods [3]. A survey on automatic facial expression analysis can be found in [4].

Information fusion is a hot research topic in biometrics, where a multibiometric system usually achieves higher recognition rate than a single biometric one. Combining multiple biometrics modalities is highly issue investigated by researchers working in this field [5]. Other application areas had benefit less by information fusion strategy. Regardless of the application, in a system that combines different types of information the fusion can be done, basically, at three levels: feature level, matching score level, and decision level. Tough it is believed that integration on the earlier stage of such a system leads to better performance than integration at the final level; just a few works has touched this type of information fusion, due to, especially, features incompatibility. Although promising, no much work has been dedicated for employing fusion information with respect to facial expression recognition task. One remarkable work is presented in [6], where the authors developed a system which uses two types of features. The first type is the geometric positions of a set of fiducial points on a face. The second type is a set of multi-scale and multi-orientation Gabor wavelet coefficients extracted from the face image at the fiducial points. They are further used independently and jointly as the input of a two-layer perceptron. However, when combined, the features do not lead to a high improvement in the classification accuracy.

We built a system which uses two sorts of information: appearance - based features and a geometrical-based features. These features are then concatenated into a single feature vector at feature level. The appearance-based features are extracted with the help of the Discriminant Non-negative Matrix Factorization (DNMF) algorithm [7]. The geometric-based features are extracted with the help of an automatic system that segments the salient features represented by eyebrows, eyes and mouth, which are relevant for the facial expression. Further, five geometrical difference distances are computed between the geometric coordinates corresponding to neutral and other facial expressions (such as disgust, happiness, surprise).

II. Geometric based approach

Let us consider a sequence of images that contains a human face acting an expression, starting with the neutral face and ending with a facial expression. The contours of facial permanent features (eyes, mouth and brows) are automatically extracted in every frame. This is accomplished by using parametric models. Two kinds of approaches exist. First of all, a coarse localization of these features based on luminance information is extracted (valley images for example) [8]. The second approach introduces models to be related to the searched contours [9]. In this paper, the following parametric models are considered: a circle for the iris, a parabola for the lower eye boundary and Bezier curves for upper eye boundary and brows and cubic curves for the mouth. Iris contour being

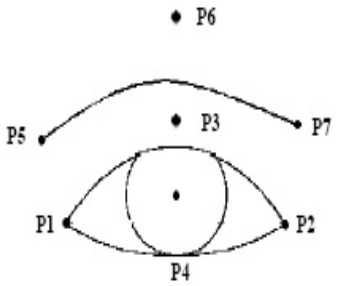

Fig. 1. Model of the right eye and eyebrow along with their keypoints.

the frontier between the dark area of iris and the eye white, it is supposed to be a circle made of points of maximum of luminance gradient. Each circle of iris maximizes [10]:

$$E = \sum_{p \in C} \overrightarrow{\nabla} I(p) \overrightarrow{n}(p) \tag{1}$$

where I is the luminance at point p, $\overrightarrow{n}(p)$ is the normal of the boundary at point p and C is a circle. Several circles scanning the search area of each iris are tested and the circle which maximizes E is selected. The radius of the circle is supposed to be known in order to reduce the computational cost. However, it is possible to test several radius values. If $A(a_1, b_1)$, $B(a_2, b_2)$, $C(a_3, b_3)$ are chosen to be three control points, the coordinates (x,y) of each point of the associated Bezier curve are defined by:

$$\begin{aligned} x &= (1-t)^2 a_1 + 2t(1-t)(a_3 - a_1) + t^2 a_2 \\ y &= (1-t)^2 b_1 + 2t(1-t)(b_3 - a_1) + t^2 b_2 \end{aligned} \tag{2}$$

Figure 1 shows the model for eye and eyebrow. In the case of the eye, for the lower boundary, a parabola is defined by points P1, P2, P4 and for the upper boundary, a Bezier curve is defined by the three control points P1, P2, P3. For the eyebrow, the usual model is generally very simple since it is a broken line defined by three points (both corners and a middle point). In this paper, we consider a Bezier curve with three control points P5, P6, P7 as the right model for eyebrows.

Eye model being defined, it is fitted on the image to be processed by the automatic extraction of keypoints (P1, P2, P3 and P4) and by the deformation of the model according to some information of maximum of luminance gradient. Indeed, the eye frontier is the limit between the eye white and the skin which is a darker area. For eyebrows segmentation P_5 and P_7 are taken into account. These two points are the corners of each eyebrow. The abscissa x_5 and x_7 of both points correspond to the left and right zero crossing of the derivative of the quantity $H(x) = \sum_{y=1}^{N_y}[255 - I(x,y)]$ and the ordinate $y_5 = y_7$ corresponds to the maximum of the quantity $V(y) = \sum_{x=1}^{N_x}[255 - I(x,y)]$, where I(x,y) is the luminance at pixel (x,y) and (N_x, N_y) represents the dimensions of the region of interest (ROI) for each eyebrow. The third control point P6 is computed using P5 and P7 as $\{x_6 = (x_5 + x_7)/2; y_6 = y_7\}$.

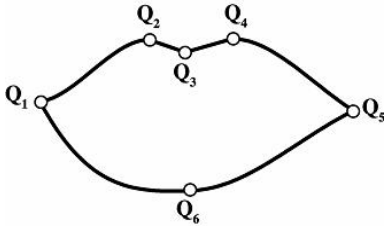

Fig. 2. Model of the mouth and its keypoints.

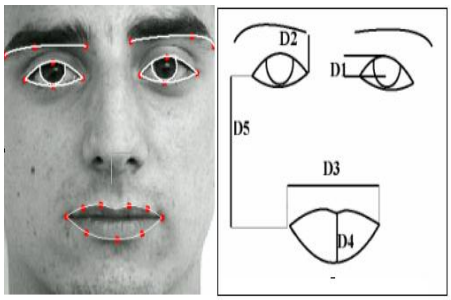

Fig. 3. Segmented face (left) and its corresponding skeleton (right).

Mouth model is more complex because mouth is a more deformable feature. We follow the model described in [10]. This approach relies on an accurate and robust quasi automatic lip segmentation algorithm. First, the upper mouth boundary and several characteristic points are detected in the first frame by using a new kind of active contour the so-called "jumping snake". Unlike classic snakes, it can be initialized far from the final edge and the adjustment of its parameters is easy and intuitive. Then, to achieve the segmentation a parametric model composed of several cubic curves is used. Its high flexibility enables accurate lip contour extraction even in the challenging case of very asymmetric mouth. Compared to existing models, it brings a significant accuracy and realism improvement. The segmentation is achieved by using an interframe tracking of the keypoints and the model parameters. The lip segmentation is depicted in Figure 2. Further details about the method can be found in [10].

Initial curves corresponding to eyes, eyebrows and mouth are deformed in order to fit the boundaries on the face image to be processed. These deformations are controlled by the maximization of a gradient flow of luminance and or chrominance through the current contour. Figure 3 (left) gives an example of facial feature segmentation.

Finally, on the resulting facial skeleton, five characteristic distances are estimated as described in Figure 3 (right): eye opening ($D1$), distance between the inner corner of the eye and the corresponding corner of the eyebrow ($D2$), mouth opening width ($D3$), mouth opening height ($D4$), distance between a mouth corner and the outer corner of the corresponding eye ($D5$)[11].

III. Appearance based approach - DNMF

Let us suppose now that we have n face images that are lexicographically scanned and stored in the columns of a $m \times n$ non-negative matrix $\mathbf{X}$. Then, each image is described by the vector $\mathbf{x}_j = [x_1, x_2, \ldots, x_m]^T$, where $j = 1, \ldots, n$ and m is the number of pixels in the image. DNMF approximates $\mathbf{X}$ (with respect to several constraints described below) by a product of two non-negative matrices (factors) $\mathbf{Z}$ and $\mathbf{H}$ of size $m \times p$ and $p \times n$, respectively, i.e. $\mathbf{X} \approx \mathbf{ZH}$. The columns of $\mathbf{Z}$ form the basis images and $\mathbf{H}$ contains in its rows the decomposition coefficients. Let us now further suppose that we have $\mathcal{Q}$ distinctive image classes and n_c is the number of image samples in a certain class $\mathcal{Q}$, $c = 1, \ldots, \mathcal{Q}$. Each image from the image database corresponds to one column of matrix $\mathbf{X}$ and belongs to one of these classes. Therefore, each column of the $p \times n$ matrix $\mathbf{H}$ can be considered as an image representation coefficient vector $\mathbf{h}_{(c)l}$, where $c = 1, \ldots, \mathcal{Q}$ and $l = 1, \ldots, n_{(c)}$. The total number of coefficient vectors is $n = \sum_{c=1}^{\mathcal{Q}} n_{(c)}$. We denote the mean coefficient vector of class c by $\boldsymbol{\mu}_{(c)} = \frac{1}{n_c} \sum_{l=1}^{n_{(c)}} \mathbf{h}_{(c)l}$ and the global mean coefficient vector by $\boldsymbol{\mu} = \frac{1}{n} \sum_{c=1}^{\mathcal{Q}} \sum_{l=1}^{n_{(c)}} \mathbf{h}_{(c)l}$. If we express the within-class scatter matrix by $\mathbf{S}_w = \sum_{c=1}^{\mathcal{Q}} \sum_{l=1}^{n_{(c)}} (\mathbf{h}_{(c)l} - \boldsymbol{\mu}_{(c)})(\mathbf{h}_{(c)l} - \boldsymbol{\mu}_{(c)})^T$ and the between-class scatter matrix by $\mathbf{S}_b = \sum_{c=1}^{\mathcal{Q}} (\boldsymbol{\mu}_{(c)} - \boldsymbol{\mu})(\boldsymbol{\mu}_{(c)} - \boldsymbol{\mu})^T$, the cost function $\mathcal{L}_{DNMF}$ associated with DNMF algorithm is written as [7]:

$$\mathcal{L}_{DNMF} = KL(\mathbf{X}||\mathbf{ZH}) + \alpha \sum_{i,j} u_{ij} - \beta \sum_i v_{ii} + \\ + \gamma \mathbf{S}_w - \delta \mathbf{S}_b, \tag{3}$$

subject to $\mathbf{Z}, \mathbf{H} \geq 0$. Here $\mathbf{U} = \mathbf{Z}^T\mathbf{Z}$, $\mathbf{V} = \mathbf{H}\mathbf{H}^T$, α, β, γ and δ are constants. The other terms appearing in the cost function have the following meaning. The first term $KL(\mathbf{X}||\mathbf{ZH}) = \sum_{i,j} \left(x_{ij} \ln \frac{x_{ij}}{\sum_k z_{ik} h_{kj}} + \sum_k z_{ik} h_{kj} - x_{ij} \right)$ is the Kullback-Leibler divergence ($k = 1, \ldots, p$) and ensures that the product $\mathbf{ZH}$ approximates as much as possible the original data $\mathbf{X}$. The second term can be further split in two as $\sum_{i,j} u_{ij} = \sum_{i \neq j} u_{ij} + \sum_i u_{ii}$, where the minimization of the first sum forces the columns of $\mathbf{Z}$ to be orthogonal in order to reduce the redundancy between basis images, while the minimization of the second term guarantees the generation of sparse features in the basis images. The third term $\sum_i v_{ii}$ aims at maximizing the total "energy" on each retained component.

Starting at iteration $t = 0$ with random positive matrices $\mathbf{Z}$ and $\mathbf{H}$, the algorithm updates their values according to an Expectation-Maximization (EM) approach, leading to the following updating rules for each iteration $t > 0$ [7]:

$$(i) \quad h_{kl(c)}^{(t)} = \frac{2\mu_c - 1}{4\xi} + \\ + \frac{\sqrt{(1 - 2\mu_c)^2 + 8\xi h_{kl(c)}^{(t-1)} \sum_i z_{ki}^{(t)} \frac{x_{ij}}{\sum_k z_{ik}^{(t)} h_{kl(c)}^{(t-1)}}}}{4\xi} \tag{4}$$

The elements h_{kl} are then concatenated for all $\mathcal{Q}$ classes as:

$$(ii) \qquad h_{kj}^{(t)} = [h_{kl(1)}^{(t)} \,|\, h_{kl(2)}^{(t)} \,|\, \ldots \,|\, h_{kl(\mathcal{Q})}^{(t)}] \tag{5}$$

where "|" denotes concatenation. The basis images are updated as:

$$(iii) \qquad z_{ik}^{(t)} = \frac{z_{ik}^{(t-1)} \sum_j \frac{x_{ij}}{\sum_k z_{ik}^{(t-1)} h_{kj}^{(t)}} h_{jk}^{(t)}}{\sum_j h_{kj}^{(t)}} \tag{6}$$

$$(iv) \qquad z_{ik}^{(t)} = \frac{z_{ik}^{(t)}}{\sum_i z_{ik}^{(t)}}, \quad \text{for all } k \tag{7}$$

The final $\mathbf{Z}$ and $\mathbf{H}$ found at the last iteration are such that $\mathbf{X} \approx \mathbf{ZH}$ as good as possible, $\mathbf{S}_w$ is as small as possible and $\mathbf{S}_b$ is as large as possible.

IV. Features and information fusion

For each subject and expression, the five distances $\{D1, D2, D3, D4, D5\}$ are stored in a 5-dimensional geometrical feature vector $\mathbf{g}$. The distances are computed for each expression with respect with the neutral pose. Therefore, in the general case where all six universal expressions are available, we have five classes for the geometrical feature vector.

In the case of DNMF approach, each image $\mathbf{x}$ is projected into the pseudoinverse of the basis images learned by DNMF. Denoting the appearance-based feature vector with $\mathbf{f}$, its expression is formed as $\mathbf{f} = \mathbf{Z}^+\mathbf{x}$, where "+" denotes matrix pseudoinversion. Notice the size of the appearance-based feature vector $\mathbf{f}$ is $p \times 1$.

Information fusion here, simply consists in concatenating both feature vectors, resulting a new $(p + 5)$ - dimensional vector $\mathbf{fg}$.

V. Experimental results and discussions

The experiments were performed by using two facial expression databases. Hammal - Caplier database [11] was used as training data for DNMF and geometric distance approach. The facial images used for testing came from the Cohn-Kanade AU-coded facial expression database [12]. The database was originally created for Action Units (AU) representation appearing in the FACS coding system and not for explicit facial expression recognition. Prior to use the samples for testing, the facial action (action units) have been converted into emotion class labels according to [3]. Hammal - Caplier database was recorded using regular (non-actor) people. Due to the difficulty for a non actor to simulate all the six universal emotions, only four expressions (*joy, disgust, surprise* and *neutral*) were used in creating the database. Therefore, in our case, the geometrical-based feature vectors form only three classes. The total number of samples corresponding to the Hammal - Caplier and Cohn-Kanade database is 192 and 104, respectively, including the neutral pose. By eliminating the neutral pose, a number of 144 and 73 subjects are attained for the aforementioned databases. Let us now denote by $\mathbf{F}_{train}$, $\mathbf{G}_{train}$, and $\mathbf{FG}_{train}$ the matrix containing all geometric-based feature vectors, appearance-based feature vectors and

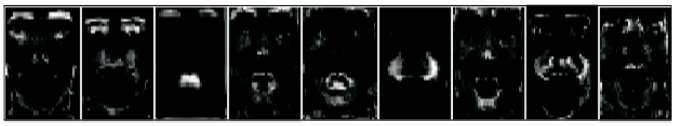

Fig. 4. Nine basis images generated by the DNMF algorithm.

the mixed features vectors corresponding to the training images. Also, denote by $\mathbf{F}_{test}$, $\mathbf{G}_{test}$, and $\mathbf{FG}_{test}$ the matrix containing all geometric-based feature vectors, appearance-based feature vectors and the mixed features vectors corresponding to the test images.

The experiments were undertaken in the following four scenarios: a) classification of the all four expressions by using DNMF algorithm, b) classification of the three expressions (without neutral) using only features extracted by DNMF, c) classification of the three expressions when using only geometrical-based features, and d) classification of three expressions by using mixed features. As mentioned in the previous Section, the reason for which the neutral was eliminated is that the geometrical-based feature extraction is based on computing five geometrical distances between each frame containing one expression and the neutral state. The number of basis images retrieved by the DNMF were varying from the set $p \in \{9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169\}$. We used kNN and kmeans as classifiers. The number of neighbors for kNN were selected from $k \in \{1, 2 \ldots, 15\}$. Table I presents the results corresponding to the lowest p, and k necessary for achieving the maximum accuracy in the case of the four scenarios.

TABLE I

CLASSIFICATION ACCURACY (%) FOR KNN AND KMEANS CLASSIFIERS. HERE p REFERS TO THE NUMBER OF BASIS IMAGES AND k REPRESENT THE MINIMUM NUMBER OF NEIGHBORS CORRESPONDING TO THE MAXIMUM ACCURACY ACHIEVED BY kNN CLASSIFIER.

	kNN			$kmeans$	
	$accuracy$	k	p	$accuracy$	p
a)	77.88	3	25	66.34	9
b)	86.30	9	9	87.67	16
c)	69.86	11	-	73.97	-
d)	90.41	7	9	91.78	9

When only DNMF features are taken into account (case a)), the maximum accuracy of 77.88% is obtained by kNN classifier having 3 neighbors and for 25 basis images. We must notice that DNMF has been also applied for facial expression classification in the case of Cohn-Kanade database in [7]. However, there, the same database was used for both training and testing (which is a standard procedure), while here, different databases were used for training and testing. Therefore, the accuracy in this case is lower than the one when the same database is used. Discarding the neutral expression, the classification accuracy increased for both classifiers (case b)). As it can be seen from the Table, the accuracy corresponding to the geometrical features is much lower compared to the previous case. Moreover, eleven neighbors are necessary for the kNN to reach the maximum accuracy. Last row shows the performance yielded by both classifiers for the fused feature vector. The results reveal that, compared with the second best case (b)), the accuracy increased from 86.30% to 90.41% for kNN and from 87.67% to 91.78% for $kmeans$ classifier. Another issue that can be noticed is related to the minimum number p of basis images necessary to achieve a maximum accuracy. As can be seen only 9 DNMF basis images were sufficient for achieving the highest recognition accuracy, which makes this algorithm attractive from the storage view point (especially when the dimension of the images is high).

VI. CONCLUSION

In this paper, we have attempted to improve the accuracy of a single facial expression recognition system by applying information fusion of two types of features. As experimental results shows, the idea of combining features has leaded to a more accurate facial expression recognition system.

ACKNOWLEDGMENT

This work has been conducted in conjunction with the "SIMILAR" European Network of Excellence on Multimodal Interfaces of the IST Programme of the European Union (www.similar.cc).

REFERENCES

[1] Y. Kenmochi Y. Shinza, Y. Saito and K. Kotani, "Facial expression analysis by integrating information of feature-point positions and gray levels of facial images," *Proc. IEEE Proc. Int. Conf. on Image Processing*, 2000.

[2] T. Kanade Y. Tian and J.Cohn, "Recognition actions units for facial expression analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[3] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, 2000.

[4] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 1, no. 30, pp. 259–275, 2003.

[5] A. Ross A. K. Jain and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 1, pp. 4–20, 2004.

[6] M. Schuster Z. Zhang, M. Lyons and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," *Proc. of Third IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 454–459, 1998.

[7] I. Buciu and I. Pitas, "A new sparse image representation algorithm applied to facial expression recognition," *Proc. IEEE Workshop on Machine Learning for Signal Processing*, pp. 539–548, 2004.

[8] K. Sobottka and I. Pitas, "Looking for faces and facial features in color images," *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, Russian Academy of Sciences*, vol. 7, no. 1, pp. 124–137, 1997.

[9] T. Kanade Y. Tian and J.Cohn, "Dual state parametric eye tracking," *Proc. of the 4th Int. Conf. on Automatic Face and Gesture Recognition*, pp. 110–115, 2000.

[10] A. Caplier Z. Hammal, N. Eveno and P-Y Coulon, "Parametric models for facial features segmentation," *Signal Processing*, 2006.

[11] M. Rombaut Z. Hammal, A. Caplier, "A fusion process based on belief theory for classification of facial basic emotions," *Proc. Fusion'2005 the 8th International Conference onInformation fusion (ISIF 2005)*, 2005.

[12] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *Proc. Fourth IEEE Int. Conf. Face and Gesture Recognition*, pp. 46–53, March 2000.