REAL TIME FACIAL EXPRESSION RECOGNITION FROM IMAGE SEQUENCES USING SUPPORT VECTOR MACHINES

I. Kotsia and I. Pitas

Aristotle University of Thessaloniki
Department of Informatics
Box 451
54124 Thessaloniki, Greece

ABSTRACT

In this paper, two novel real-time methods are proposed for facial expression recognition in image sequences. The user manually places some of the Candide grid's points to the face depicted at the first frame. The grid adaptation system tracks the entire grid as the facial expression evolves through time, thus producing a grid that corresponds to the greatest intensity of the facial expression, as shown at the last frame. Certain points that are involved into creating the Facial Action Units (FAUs) movements are selected. Their geometrical displacement information, defined as the coordinates' difference between the last and the first frame, is extracted to be the input to a bank of Support Vector Machine (SVM) classifiers that are used to recognize either the six basic facial expressions or eight chosen FAUs. The results show a recognition accuracy of approximately 98% and 94% for direct and FAU based facial expression recognition, respectively.

1. INTRODUCTION

Several research efforts have been done regarding facial expression recognition during the past two decades, due to its importance for human centered interfaces. The facial expressions under examination were defined as a set of six basic facial expressions (anger, disgust, fear, happiness, sadness and surprise), whose combinations produce every other "complex" facial expression [1]. In order to make the recognition procedure more standardized, a set of muscle movements (known as Action Units) that produce each facial expression, was created by psychologists, thus forming the so called $Facial \quad Action \quad Coding \quad System \quad (FACS)$ [2].

A survey on automatic facial expression recognition can be found in [3], [4] and [5].

In the current paper, two novel fast methods for recognizing dynamic facial expressions either directly or by detecting first the Facial Action Units (FAUs) are proposed, that use Support Vector Machines (SVM) classifiers. The user has to manually place in the beginning some of the Candide grid's points to a face depicted at the first frame of the image sequence under examination. The tracking system follows the facial expression evolving through time to reach its highest intensity, producing at the same time the grid that corresponds to it. A subset of the Candide grid's point is chosen, as the one that is responsible for the formation of movement as described by the FAUs. The geometrical displacement of those points, defined as the difference of each point's coordinates between the last and the first frame of the image sequence, are used as an input to a bank of SVMs. The experiments performed using the Cohn-Kanade database indicate a recognition accuracy of 97.75% or 93.7% when recognizing six basic facial expressions using the direct approach or the FAU-based approach, respectively.

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).

2. SYSTEM DESCRIPTION

The diagram of the system used for the experiments is shown in Figure 1. The system is composed of two subsystems, one for geometrical information extraction and one for geometrical information classification.

Facial expressions can be described as combinations of Facial Action Units (FAUs), as proposed by [6]. As can be seen at the second column of Table 2, the FAUs that are necessary for fully describing all facial expressions according to the Facial Action Coding System (FACS), are the 17 FAUs 1, 2, 4, 5, 6, 7, 9, 10, 12, 15, 16, 17, 20, 23, 24, 25 and 26. A subset of FAUs is chosen (FAUs 5, 9, 12, 15, 16, 20, 23 and 24) as those that appear once or twice in the whole set of facial expressions (shown at the third column of Table 2).

2.1. Geometrical displacement information extraction

The geometrical information extraction is done by a grid adaptation system, based on deformable models. The user

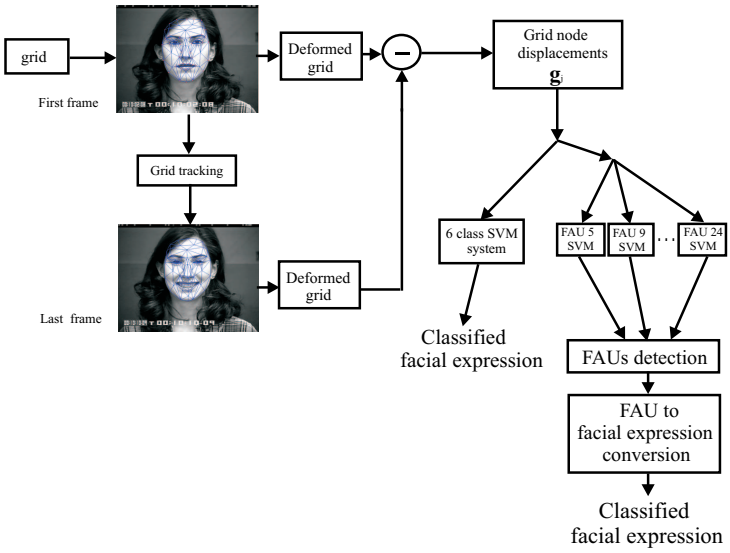

Fig. 1. System description

Expression	FAUs coded description [6]	FAUs chosen
Anger	$4+7+$ $+(((23 \text{ or } 24) \text{ with or not } 17)$ or $(16+(25 \text{ or } 26))$ or $(10+16+(25 \text{ or } 26)))$ with or not 2	23 or 24
Disgust	$((10 \text{ with or not } 17) \text{ or } (9 \text{ with or not } 17))+(25 \text{ or } 26)$	9
Fear	$(1+4)+(5+7)+20+$ $+(25 \text{ or } 26)$	20
Happiness	$6+12+16+$ $+(25 \text{ or } 26)$	$12+16$
Sadness	$1+4+(6 \text{ or } 7)+$ $15+17+(25 \text{ or } 26)$	15
Surprise	$(1+2)+(5 \text{ without } 7)+26$	5

Table 1. The FAUS chosen to be used for the SVMs facial expression recognition system

has to manually place some of Candide grid points to the face depicted at the first frame of the image sequence. The points around the eyes, eyebrows and mouth are the ones with the greatest importance. The software automatically adjusts the grid to the face and then tracks it through the image sequence, as it evolves through time [7]. At the end, the grid adaptation software produces the deformed Candide grid that corresponds to the facial expression with the greatest intensity.

The deformed Candide grid produced by the grid adaptation software, is constructed by 104 points. A subset of 62 points are chosen, as those that control the movement described by the 17 FAUs used for describing facial expressions (shown in figure 1).

The classification is performed based only in geometrical information, without taking into consideration any luminance or color information.

In the case of direct facial expression recognition, let $\mathcal{U}$ be the video database that contains the facial videos clips, that are clustered into 6 different classes $\mathcal{U}_k,\ k=1,\ldots,6$, each one representing one of 6 basic facial expressions (anger, disgust, fear, happiness, sadness and surprise).

In the case of FAU-based facial expression recognition, for every FAU the database is clustered into 2 different classes $\Theta_i^k,\ i=1,2$ for the k-th FAU ($k=\{1,\ldots,8\}$). The first class, Θ_1^k, represents the presence of the FAU under examination (each one of FAUs 5, 9, 12, 15, 16, 20, 23 and 24) at the PFEG being processed, while the second one, Θ_2^k, represents its absence.

The geometrical information used is the displacement of one point $\mathbf{d}_j^i$, defined as the difference between the last and the first frame's coordinates:

$$\mathbf{d}_j^i = \begin{bmatrix} \Delta x_j^i \\ \Delta y_j^i \end{bmatrix}, \quad i \in \{1,\ldots,K\} \quad \text{and} \quad j \in \{1,\ldots,N\} \tag{1}$$

where i is the number of points taken under consideration, here K, equal to 62 and j is the the number of image sequences to be examined.

For every facial video in the training set, a feature vector $\mathbf{g}_j$ is created, containing the geometrical displacement of every grid node:

$$\mathbf{g}_j = \begin{bmatrix} \mathbf{d}_j^1 \\ \mathbf{d}_j^2 \\ \vdots \\ \mathbf{d}_j^K \end{bmatrix} \tag{2}$$

having $F = 62 \cdot 2 = 124$ dimensions.

2.2. Geometrical displacement information classification

In the case of direct facial expression recognition, the feature vector $\mathbf{g}_j \in \Re^F$ is used as an input to a multi class SVM, labelled properly with the true corresponding facial expression. The output of the SVM system is a label that classifies the grid under examination to one of the six basic facial expressions.

In the case of FAU-based facial expression recognition, for each FAU under examination, the feature vector $\mathbf{g}_j \in \Re^F$ is used as an input, labelled properly with the true corresponding label l_j. The output of each SVM classifier (8 two class SVMs in total)is a label that specifies if the specific FAU is activated ($l_j = 1$) or not ($l_j = -1$).

3. SUPPORT VECTOR MACHINES

A brief introduction to the multiclass SVM theory will be outlined below. The interested reader can refer to [9] and the references therein for formulating and solving multiclass SVM optimization problems.

Suppose the training data:

$$(\mathbf{g}_1, l_1), \ldots, (\mathbf{g}_N, l_N) \tag{3}$$

where $\mathbf{g}_j \in \Re^F \quad j = 1, \ldots, N$ the deformation feature vectors and $l_j \in \{1, \ldots, 6\} \quad j = 1, \ldots, N$ are the facial expression labels of the feature vector. The approach implemented for multiclass problems used for direct facial expression recognition is the one proposed in [9] that solves only one optimization problem. It constructs 6 (six facial expressions) two-class rules where the $k-$th function $\mathbf{w}_k^T \phi(\mathbf{g}_j) + b_k$ separates training vectors of the class k from the rest of the vectors. Hence, there are 6 decision functions, all obtained by solving one SVM problem. The formulation is as follows:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{k=1}^{6} \mathbf{w}_k^T \mathbf{w}_k + C \sum_{j=1}^{N} \sum_{k \neq l_j} \xi_j^k \\ \mathbf{w}_{l_j}^T \phi(\mathbf{g}_j) + b_{l_j} & \geq \mathbf{w}_k^T \phi(\mathbf{g}_j) + b_k + 2 - \xi_j^k \\ \xi_j^k \geq 0, \quad & j = 1, \ldots, N, \quad k \in \{1, \ldots, 6\} \backslash l_j. \end{aligned} \tag{4}$$

where ϕ is the function that maps the deformation vectors to a higher dimensional space, where the data are supposed to be linearly or near linearly separable. C is the penalty parameter for non linear separability. The vector $\mathbf{b} = [b_1 \ldots b_6]^T$ is the bias vector and $\boldsymbol{\xi} = [\ldots, \xi_i^m, \ldots]^T$ is the slack variable vector. Then the decision function is:

$$h(\mathbf{g}) = \underset{k=1,\ldots,6}{\operatorname{argmax}} (\mathbf{w}_k^T \phi(\mathbf{g}) + b_k). \tag{5}$$

For the creation of a two-class SVM classifier, used in the FAU-based facial expression recognition approach, in order to train the SVM network, the following minimization problem should be solved [9]:

$$\begin{aligned} \min_{\mathbf{w}_k, b_k, \boldsymbol{\xi}^k} \quad & \frac{1}{2}\mathbf{w}_k^T \mathbf{w}_k + C_k \sum_{j=1}^{N} \xi_j^k \\ \mathbf{w}_k^T \phi(\mathbf{g}_j) + b_k & \geq 1 - \xi_j^k, \quad \text{if } y_j = 1 \\ \mathbf{w}_k^T \phi(\mathbf{g}_j) + b_k & \leq -1 + \xi_j^k, \quad \text{if } y_j = -1 \\ \xi_j^k \geq 0, \quad & j = 1, \ldots, N \end{aligned} \tag{6}$$

where b_k is the bias for the k-th SVM, $\boldsymbol{\xi}^k = [\ldots, \xi_i^k, \ldots]$ is the slack variable vector and C_k is the penalty. After solving (6), the function that decides whether the k-th FAU is activated by a test displacement feature vector $\mathbf{g}$ is:

$$f_k(\mathbf{g}) = \mathbf{w}_k^T \phi(\mathbf{g}) + b_k. \tag{7}$$

In this formulation, a nonlinear mapping ϕ has been used for a high dimensional feature mapping. This mapping is defined by a positive kernel function, $k(\mathbf{g}_m, \mathbf{g}_n)$, specifying an inner product in the feature space and satisfying the Mercer condition [9]:

$$\phi(\mathbf{g}_m)^T \cdot \phi(\mathbf{g}_n) = k(\mathbf{g}_m, \mathbf{g}_n). \tag{8}$$

Many functions can be used as the kernel of the SVM system. The most common ones that are also used for our experiments are the d degree polynomial function:

$$k({\mathbf{g}_m}^T, \mathbf{g}_n) = ({\mathbf{g}_m}^T \cdot \mathbf{g}_n + 1)^d \tag{9}$$

and the Radial Basis Function (RBF) kernel:

$$k(\mathbf{g}_m, \mathbf{g}_n) = \exp(-\frac{\| \mathbf{g}_m - \mathbf{g}_n \|^2}{2\sigma^2}). \tag{10}$$

4. EXPERIMENTAL RESULTS

The database used for the experiments was the Cohn-Kanade database [2], which is encoded into combinations of Action Units. These combinations were translated into facial expressions according to [6]. For each person, the image sequence was created and processed by the grid adaptation system, based on deformable models. In figure 2, a sample of image sequences of one person from the database used for the experiments, is shown. The experiments indicated that the whole system is fast enough to fulfill a real-time system's requirements, since it is able to process 20 frames per second. The classification accuracy was measured as the percentage of the correctly classified facial expressions.

In the case of direct facial expression recognition, the leave-one-out method was used. Therefore, the database consisted of 222 image sequences and the polynomial function used for the creation of the polynomial kernel, was of degree 3. The accuracy achieved was equal to 97,75% when the 6 basic facial expressions were under examination.

The confusion matrix [8] has been computed. It is a $n \times n$ matrix containing the information about the actual class label l_j, $j = 1, .., n$ (in its rows) and the label obtained through classification p_j, $j = 1, .., n$ ones (in its columns). The diagonal entries of the confusion matrix are the number of facial expressions that are correctly classified, while the off-diagonal entries correspond to misclassification. The confusion matrix showed that the ambiguous facial expression was anger, since it was the only one misclassified as another one of the remaining 5 basic facial expressions (mostly misclassified as sadness and then as disgust).

The abbreviations an, di, fe, ha, sa and su represent anger, disgust, fear, happiness, sadness and surprise respectively, and lab_{ac}, lab_{clas} represent the actual and the classified label of the video sequence, respectively.

Table 2. Confusion matrix for dynamic direct facial expression recognition.

$lab_{ac} \backslash lab_{cl}$	an	di	fe	ha	sa	su
an	32	0	0	0	0	0
di	1	37	0	0	0	0
fe	0	0	37	0	0	0
ha	0	0	0	37	0	0
sa	4	0	0	0	37	0
su	0	0	0	0	0	37

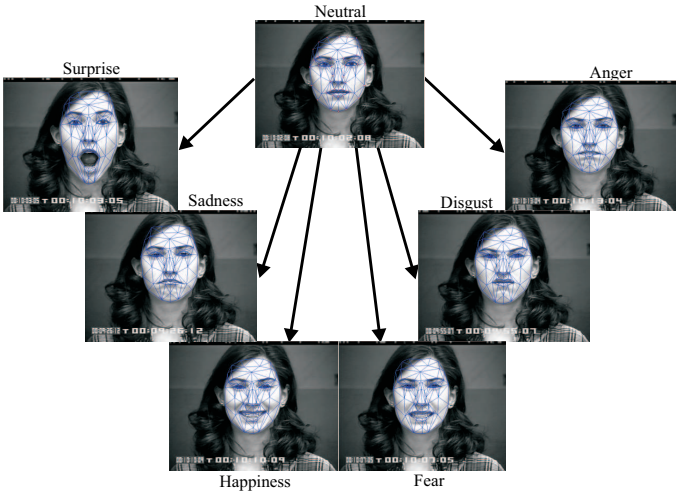

Fig. 2. An example of each facial expression for a poser from the Cohn-Kanade database.

In the case of FAU-based facial expression recognition, a total of 365 image sequences were used. Approximately the 87% of the image sequences were used for the training process and the rest (13%) for the testing process. The radial basis function used for the creation of the kernel had a σ equal to $\frac{\sqrt{2}}{2}$. The accuracy achieved was equal to 93.7% when 8 basic FAUs were under examination.

5. CONCLUSION

Two novel fast methods for direct and FAU based facial expression recognition are proposed in this paper. The user initializes some of the Candide grid nodes on the facial image depicted at the first frame of the image sequence. The Candide nodes that influence the formation of FAUs are used in our system. The tracking system used, based on deformable models, tracks the facial expression as it evolves over time, by deforming the Candide grid eventually producing the grid that corresponds to the facial expression greatest intensity typically depicted at the last facial video frame. Their geometrical displacement, defined as their coordinate difference between the last and the first frame, is used as an input to the SVM system. In the case of direct facial expression recognition, this system is composed of one six-class SVMs, one for each one of the 6 basic facial expressions (anger, disgust, fear, happiness, sadness and surprise) to be recognized. When FAU-based facial expression recognition is attempted, the SVM system consists of 8 one-class SVMs, one for each one of the 8 chosen FAUs used. The proposed methods, achieve a facial expression recognition accuracy of 97,75% and 93,7% respectively. The achieved accuracy is better than any other reported in the literature so far for the Cohn-Kanade database, at least according to the authors knowledge.

6. REFERENCES

[1] P. Ekman, and W.V. Friesen, "Emotion in the Human Face," *Prentice Hall*, 1975.

[2] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis," *Proceedings of IEEE International Conference on Face and Gesture Recognition*, 2000.

[3] M. Pantic, and L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[4] B. Fasel, and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *Pattern Recognition*, 2003.

[5] Y.Tian, T.Kanade and J.Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Transactions of Pattern, Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97-115, 2001.

[6] M. Pantic, and L. J. M. Rothkrantz, "Expert System for Automatic Analysis of Facial Expressions," *Image and Vision Computing*, 2000.

[7] S. Krinidis, and I. Pitas, "Statistical Analysis of Facial Expressions for Facial Expression Synthesis," *submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[8] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic Classification of Single Facial Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1999.

[9] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415-425, 2002.