MULTICLASS SUPPORT VECTOR MACHINES AND METRIC MULTIDIMENSIONAL SCALING FOR FACIAL EXPRESSION RECOGNITION

Irene Kotsia[†], *Stefanos Zafeiriou*[†], *Nikolaos Nikolaidis*[†] *and Ioannis Pitas*[†]

[†]Aristotle University of Thessaloniki, Department of Informatics
Thessaloniki, Greece
email: {*ekotsia, dralbert, nikolaid, pitas*}*@aiia.csd.auth.gr*

ABSTRACT

In this paper, a novel method for the recognition of facial expressions in videos is proposed. The system first extracts the deformed Candide facial grid that corresponds to the facial expression depicted in the video sequence. The mean Euclidean distance of the deformed grids is then calculated to create a new metric multidimensional scaling. The classification of the sample under examination to one of the 7 possible classes of facial expressions, i.e. anger, disgust, fear, happiness, sadness, surprise and neutral, is performed using multiclass SVMs defined in the new space. The experiments were performed using the Cohn-Kanade database and the results show that the above mentioned system can achieve an accuracy of 95.6%.

1. INTRODUCTION

In the last two decades, facial expression recognition has attracted scientific interest due to its vital role in many applications such as human centered interfaces, e.g. virtual reality, video-conferencing, user profiling and customer satisfaction studies for broadcast and web services. Psychologists have defined a set of facial expressions that are thought to be expressed in a similar way all over the world, thus making the facial expression recognition more standard. These facial expressions are anger, disgust, fear, happiness, sadness and surprise [1]. These basic facial expressions in addition with the neutral state are the target of facial expression recognition systems developed nowadays. A survey on automatic facial expression recognition can be found in [2].

Recently, in [3], a method for the recognition of the six basic facial expressions has been proposed. The user manually places some of the Candide grid's points to the face depicted at the first frame. A grid adaptation system, based on deformable models, tracks the entire Candide grid as the facial expression evolves through time, thus producing a grid that corresponds to the greatest intensity of the facial expression, as depicted at the last frame. The geometrical displacement information of the grid points, defined as the coordinates' difference between the last and the first frame, is extracted to be the input to a six class SVMs system. The system requires the presence of neutral state in order to calculate the geometrical information that is used for classification. Thus, the recognition of neutral state is not feasible.

In this paper, a novel method for the recognition of facial expressions is proposed. The system incorporates the same tracking system as in [3] and the deformed grid that corresponds to the facial expression depicted on the last frame is obtained. Unlike the method proposed in [3], knowledge of the grid that corresponds to the neutral state is not necessary, as the proposed system requires only the deformed grid obtained from the grid tracking system and does not need to calculate the grid coordinates difference between the neutral and fully expressed image. Thus, the system can take as an input a video sequence starting from any facial expression and classify each frame to one of the seven facial expression classes (6 basic facial expressions plus neutral state). The system achieved an accuracy rate of 95.6% on experiments performed in the Cohn-Kanade database.

2. SYSTEM DESCRIPTION

The diagram of the system used for the experiments is shown in Figure 1. The information extraction subsystem consists of the grid tracking system described in [4]. The extracted information, is used as an input to the information processing subsystem, that includes the calculation of the mean Euclidean distances and the embedding part. Finally, the information classification subsystem consists of a 7-class SVMs system that classifies the embedded deformed grid into one of the 7 facial expression classes under examination.

This work was supported by the "SIMILAR" European Network of Excellence on Multimodal Interfaces of the IST Programme of the European Union (www.similar.cc) for Ms. Kotsia and by project 03ED849 co-funded by the European Union and the Greek Secretariat of Research and Technology (Hellenic Ministry of Development) of the Operational Program for Competitiveness within the 3rd Community Support Framework for Mr. Zafeiriou.

1-4244-1566-7/07/$25.00 ©2007 IEEE.

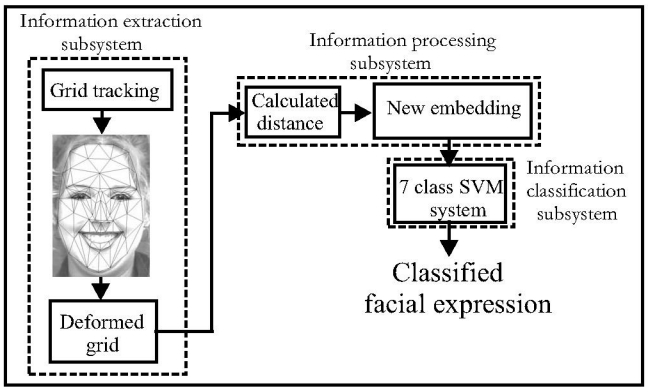

Fig. 1. Flow chart of the proposed system

3. INFORMATION EXTRACTION SUBSYSTEM

The Candide grid [5] that was used is a parameterized face mask specifically developed for model-based coding of human faces. A frontal view of the model can be seen in Figure 2. The low number of its triangles allows fast face animation with moderate computing power.

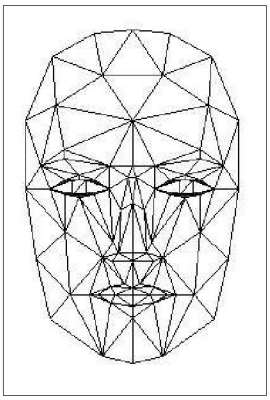

Fig. 2. The Candide grid.

The geometrical information extraction is performed by a grid tracking system, based on deformable models [4], that uses a pyramidal implementation of the well-known Kanade-Lucas-Tomasi (KLT) algorithm. The user has to place manually a number of Candide grid nodes on the corresponding positions of the face depicted at the first frame of the image sequence. For automatic grid displacement elastic graph matching techniques [6] can be used. The algorithm automatically adjusts the grid to the face and then tracks it through the image sequence, as it evolves through time to reach its highest intensity, thus producing the deformed Candide grid.

The deformed Candide grids are firstly normalized. The normalization procedure involves their scaling, translation and rotation. More precisely, the scaling was performed in such a way that the width and the height of all deformed grids would be the same for all test samples. The translation included the translation of each deformed grid so that the node corresponding to the tip of the nose would be placed at the center of the coordinate system. Last, the rotation included the rotation of each deformed grid so that the grids that were produced would be aligned. In order to achieve that, an angle was defined as the one created by the vertical line that connects the center of the mouth with the center of the forehand and by the horizontal line that connects the inner nodes of the eyes.

4. METRIC MULTIDIMENSIONAL SCALING

4.1. Mean Grid Node Euclidean distance

Given two Candide grid sets of points: $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_p\}$ and $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ the distance that is adopted is the mean Euclidean distance, defined as:

$$d_M(\mathcal{A}, \mathcal{B}) = \frac{1}{N(\mathcal{A})} \sum \|\mathbf{a}_i - \mathbf{b}_i\|. \quad (1)$$

This distance is similar to the Euclidean distance between two grids but is calculated in a node-wise manner as we are interested in the deformation of each node independently. It should be noted here, that for the Candide grid the nodes correspondences are known. The set of points creating the grid is combined with a predefined way, thus the correspondences of points between two random Candide grids are known, i.e. point $\mathbf{a}_i$ corresponds to point $\mathbf{b}_i$. In the proposed approach we use the mean Euclidean distance in (1) in order to create a feature space, using a new embedding so as to define later a multiclass SVM classifier in this space.

4.2. Embedding to the new space

It can be easily proven that the measure in (1) satisfies the following properties:

- reflectivity i.e., $d_{MH}(\mathcal{A}_i, \mathcal{A}_i) = 0$
- positivity i.e., $d_{MH}(\mathcal{A}_i, \mathcal{A}_j) > 0$ if $\mathcal{A}_i \neq \mathcal{A}_j$
- symmetry i.e., $d_{MH}(\mathcal{A}_i, \mathcal{A}_j) = d(\mathcal{A}_j, \mathcal{A}_i)$
- triangle inequality i.e., $d(\mathcal{A}_i, \mathcal{A}_j) \leq d(\mathcal{A}_i, \mathcal{B}) + d(\mathcal{B}, \mathcal{A}_j)$ $\forall \mathcal{A}_i, \mathcal{A}_j, \mathcal{B}$ Candide grid points sets.

Thus, the mean Euclidean distance used is a metric measure [7]. We will use this metric in order to define an embedding in a new multidimensional space [8] [9]. The procedure that will be described below is like a Principal Component Analysis [10], but is now applied in a node wise manner

(equation 1) since we are interested in the deformation of each node independently.

Let $\{\mathcal{A}_1, \ldots, \mathcal{A}_N\}$ be the set of training facial grid database. The dissimilarity matrix of the training is defined as:

$$[\mathbf{D}]_{i,j} = d_M(\mathcal{A}_i, \mathcal{A}_j). \tag{2}$$

We will use the similarity matrix $\mathbf{D}$ in order to define an embedding $\mathbf{X} \in \Re^{k \times N}$, where $k \leq N$ is the dimensionality of the embedding and the i-th column of $\mathbf{X}$, denoted as $\mathbf{x}_i$, corresponds to the feature vector of the facial grid $\mathcal{A}_i$ in the new space. In order to find the embedding $\mathbf{X}$, the matrix $\mathbf{B}$ is defined as:

$$\mathbf{B} = -\frac{1}{2}\mathbf{JDJ} \tag{3}$$

where $\mathbf{J} = \mathbf{I}_{N\times N} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T \in \Re^{N\times N}$ is the centering matrix, where $\mathbf{I}_{N\times N}$ is the $N \times N$ identity matrix and $\mathbf{1}_N$ is the N-dimensional vector of ones. The matrix $\mathbf{J}$ projects the data so that the embedding $\mathbf{X}$ has zero mean. The eigen-decomposition of the matrix $\mathbf{B}$ will give us the desired embedding. The matrix $\mathbf{B}$ is positive semi-definite (i.e., it has real and non-negative eigenvalues), if and only if the distance matrix $\mathbf{D}$ is Euclidean [7]. Let p be the number of positive eigenvalues of matrix $\mathbf{B}$. Then, the matrix $\mathbf{B}$ can be written as:

$$\mathbf{B} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}} \begin{bmatrix} \mathbf{I}_{p\times p} & \\ & \mathbf{0} \end{bmatrix} \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^T = \mathbf{X}^T\mathbf{M}\mathbf{X} \tag{4}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with the diagonal consisting of the p positive eigenvalues, which are presented in the following order: first, positive eigenvalues in decreasing order and then the zero values. The matrix $\mathbf{I}_{p\times p}$ is the identity $p \times p$ matrix. The matrix $\mathbf{X}_p$ is the embedding of the set of facial grids in the new space $\Re^p$ [7]:

$$\mathbf{X}_p = \mathbf{\Lambda}_p^{\frac{1}{2}}\mathbf{Q}_p^T \tag{5}$$

where $\mathbf{\Lambda}_p$ contains only the non-zero diagonal elements of $\mathbf{\Lambda}$ and $\mathbf{Q}_p$ is the matrix with the corresponding eigenvectors.

The new embedding is purely Euclidean. As already mentioned, the vector $\mathbf{x}_i^p$, i.e. the i-th column of the matrix $\mathbf{X}_p$ corresponds to the feature vector of the grid $\mathcal{A}_i$ in the new space.

5. MULTICLASS SVMS FOR CLASSSIFICATION

5.1. Training phase

The new space is purely Euclidean and a multi-class SVM is built now to classify the vectors $\mathbf{x}_i^l$ (the features of the $\mathcal{A}_i$ Candide training grid). The training data are $(\mathbf{x}_1^l, l_1), \ldots$ $\ldots, (\mathbf{x}_N^l, l_N)$ where $\mathbf{x}_i^l \in \Re^L$ are the feature vectors and $l_j \in \{1, \ldots, 7\}$ are the facial expression labels of the feature vectors. The multi-class SVMs problem solves only one optimization problem [11]. It constructs 7 facial expressions rules, where the $k-$th function $\mathbf{w}_k^T\phi(\mathbf{x}_i^l)+b_k$ separates training vectors of the class k from the rest of the vectors, by minimizing the objective function:

$$\min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi}} \quad \frac{1}{2}\sum_{k=1}^{7}\mathbf{w}_k^T\mathbf{w}_k + C\sum_{j=1}^{N}\sum_{k\neq l_j}\xi_j^k \tag{6}$$

subject to the constraints:

$$\begin{aligned} \mathbf{w}_{l_j}^T\phi(\mathbf{x}_i^l) + b_{l_j} &\geq \mathbf{w}_k^T\phi(\mathbf{x}_i^l) + b_k + 2 - \xi_j^k \\ \xi_j^k \geq 0, \quad j &= 1, \ldots, N, \quad k \in \{1, \ldots, 7\}\backslash l_j. \end{aligned} \tag{7}$$

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 term that penalizes the training errors. The vector $\mathbf{b} = [b_1 \ldots b_7]^T$ is the bias vector and $\boldsymbol{\xi} = [\xi_1^1, \ldots, \xi_i^k, \ldots, \xi_N^7]^T$ is the slack variable vector.

It is not necessary to know the explicit form of the function ϕ, since only the close form of the dot products in $\mathcal{H}$, the so called *kernel trick* is required:

$$h(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T\phi(\mathbf{y}) \tag{8}$$

where the function h is known as kernel function. The typical kernels that have been used in the literature have been the polynomial and the Radial Basis Functions (RBF) kernels:

$$\begin{aligned} h(\mathbf{x}, \mathbf{y}) &= \phi(\mathbf{x})^T\phi(\mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^d \\ h(\mathbf{x}, \mathbf{y}) &= \phi(\mathbf{x})^T\phi(\mathbf{y}) = e^{-\gamma(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} \end{aligned} \tag{9}$$

where $d \in \mathcal{N}$ is the degree of the polynomial and γ is the spread of the Gaussian cluster.

For the solution of the optimization problem (6) subject to the constraints (7) one can refer to [11, 12]. The solution of (6) subject to (7) provides us with normal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_7$ and with seven bias terms $b_1, \ldots, b_7$.

5.2. Facial grid classification using the trained SVMs

In this Section we will show how features from previously "unseen" facial grids are embedded in the new Euclidean space using the proposed similarity measure. The features are afterwards classified with the multi-class SVMs system. Let $\{\mathcal{G}_1, \ldots, \mathcal{G}_n\}$ be a set of n testing facial grids. We create the matrix $\mathbf{D}_n \in \Re^{n\times N}$, with $[\mathbf{D}_n]_{i,j} = d_{MH}(\mathcal{G}_i, \mathcal{A}_j)$. The matrix $\mathbf{D}_n$ represents the similarity, with respect to the distance, between the n test facial grids and all the training facial grids. The matrix $\mathbf{B}_n \in \Re^{n\times N}$ of inner products that relates all the new (test) facial grids to all facial grids from the training set is then found as follows:

$$\mathbf{B}_n = -\frac{1}{2}(\mathbf{D}_n\mathbf{J} - \mathbf{UDJ}) \tag{10}$$

where $\mathbf{J}$ is the centering matrix and $\mathbf{U} = \frac{1}{N}\mathbf{1}_n\mathbf{1}_N^T \in \Re^{n\times N}$. The embedding $\mathbf{Y}_n \in \Re^{l\times n}$ of the test facial grids is defined as:

$$\mathbf{Y}_n = \mathbf{\Lambda}_l^{-\frac{1}{2}}\mathbf{Q}_l^T\mathbf{B}_n^T. \quad (11)$$

The columns of the matrix $\mathbf{Y}_n$ are the features used for classification. Let $\mathbf{y}_{i,n} \in \Re^l$ be the i-th column of the matrix $\mathbf{Y}_n$, i.e. the vector that contains the features of the grid $\mathcal{G}_i$. The classification of $\mathcal{G}_i$ to one of the seven facial expression classes is performed by the decision function:

$$h(\mathcal{G}_i) = \arg\max_{k=1,\ldots,7}(\mathbf{w}_k^T\phi(\mathbf{y}_{i,n}) + b_k), \quad (12)$$

where $\mathbf{w}_k$ and b_k have been found during training, as described in Section 5.1.

6. EXPERIMENTAL RESULTS

The database used for the facial expression recognition experiments was created using the Cohn-Kanade database [13]. This database is annotated with FAUs. These combinations of FAUs were translated into facial expressions according to [2], in order to define the corresponding ground truth for the facial expressions. In Figure 3, a sample of the grids acquired for one person from the database used for the experiments, is shown. The classifier accuracy was measured

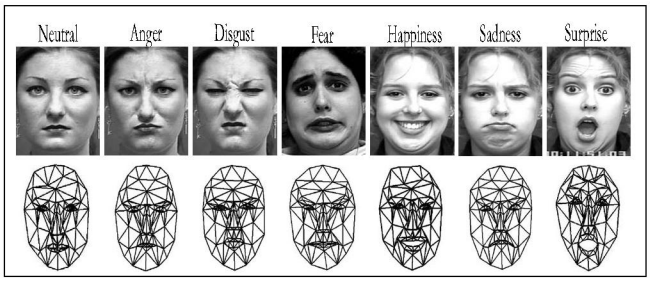

Fig. 3. An example of the grids extracted for a poser from the Cohn-Kanade database.

using the leave-one-out cross-validation approach described below, in order to make maximal use of the available data and produce averaged classification accuracy results. The image sequences contained in the database are divided into 7 classes, each one corresponding to one of the 7 facial expressions. Each class consists of the same number of fully expressive facial expression grid samples (37 facial grids for every expression). One facial expression sample from each class is used for the test set, while the remaining samples form the training set. During the training procedure the distance matrix $\mathbf{D}$ of the training samples is calculated. Afterwards, the new embedding is performed and finally the multiclass SVM system is trained.

The seven test samples are then projected in the new embedding, as described in Section 5.2, and afterwards classified using (12). Subsequently, the samples forming the test set are incorporated into the current training set and a new set of samples (one for each class) is extracted to form the new test set. The remaining samples create the new training set. This procedure is repeated until all of the samples are used as test sets. The classification accuracy is measured as the mean value of the percentages of the correctly classified facial expressions.

We have experimented with the dimensionality of the new embedding which can be modified by keeping only the p eigenvectors with the largest eigenvalues (i.e. using a matrix $\mathbf{X} \in \Re^{p\times(36\times7)}$). Figure 4 depicts the facial expression recognition rate achieved versus the dimensionality of the embedding space. The accuracy achieved with the proposed

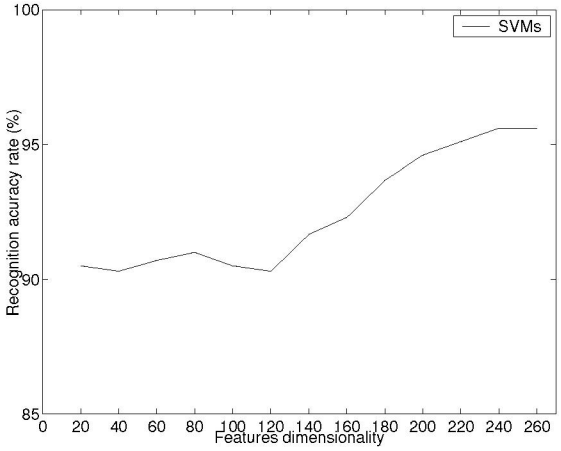

Fig. 4. Facial expression recognition rate (7 facial expressions) versus dimensionality of the new embedding in the Cohn-Kanade database

system was equal to 95.6% when SVMs were used with a polynomial kernel of degree equal to 3 for $p = 260$.

A comparison of the recognition rates achieved for each facial expression with the state of the art when six facial expressions were examined (the neutral state was not taken under consideration) can be found in [14]-[17]. As can be seen, our recognition rates are among the highest for each facial expression. A decrease of 4.1% was noticed when compared with the method proposed in [18]. This is due to the introduction of the neutral state in the recognition process. The recognition of the neutral state was attempted as almost in all cases the first image of an image sequence does not depict the neutral state. Therefore the method proposed in [18] could not be applied. Furthermore, a person may experience many psychological changes through time, thus depicting many facial expressions variations. A system should be able to recognize them and to achieve that

the recognition of the neutral state is vital.

7. CONCLUSIONS

A novel method for the classification of seven facial expressions (i.e. anger, disgust, fear, happiness, sadness, surprise and neutral) using only facial grids that have been deformed to find the facial characteristics in videos, has been presented. The mean Euclidean distance has been exploited in order to create a new embedding space and a multiclass SVM system has been defined in this space to be used for the classification of expression. Experiments showed that the proposed technique achieved an accuracy rate of 95,6% when recognizing seven facial expressions (6 basic facial expressions plus neutral).

8. REFERENCES

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

[2] 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, August 2000.

[3] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172 – 187, 2007.

[4] S. Krinidis and I. Pitas, "Statistical analysis of facial expressions for facial expression synthesis," *submitted to IEEE Transactions on Multimedia*, 2006.

[5] M. Rydfalk, "CANDIDE: A parameterized face," Tech. Rep., Linkoping University, 1978.

[6] R.P. Wurtz, "Object recognition robust under translations, deformations, and changes in background," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 769–775, July 1997.

[7] E. Pekalska, P. Paclik, and R.P.W. Duin, "A generalized kernel approach to dissimilarity-based classification," *Journal of Machine Learning Research*, vol. 2, pp. 175–211, 2001.

[8] I. Borg and P. Groenen, *Modern Multidimensional Scaling: theory and applications*, Springer-Verlag, New York, 1997.

[9] M. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, Chapman and Hall, 2001.

[10] K. Fukunaga, *Statistical Pattern Recognition*, CA: Academic, San Diego, 1990.

[11] J. Weston and C. Watkins, "Multi-class Support Vector Machines," Tech. Rep. Techical report CSD-TR-98-04, 1998.

[12] V. Vapnik, *Statistical Learning Theory*, J.Wiley, New York, 1998.

[13] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of IEEE International Conference on Face and Gesture Recognition*, March 2000, pp. 46–53.

[14] I. Cohen, N. Sebe, S. Garg, L. S. Chen, and T. S. Huanga, "Facial expression recognition from video sequences: temporal and static modelling," *Computer Vision and Image Understanding*, vol. 91, pp. 160–187, 2003.

[15] S.P. Aleksic and K.A. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multi-stream hmms," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 1, pp. 3–11, 2006.

[16] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," June 2006, vol. 8, pp. 500–508.

[17] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, June 2007.

[18] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, January 2007.