# Pairwise Facial Expression Classification

*Marios Kyperountas*[#], *Anastasios Tefas*[#], *and Ioannis Pitas*[#,*]

[#] Department of Informatics, Aristotle University of Thessaloniki, Greece
[*] Informatics and Telematics Institute, CERTH, Greece
{mkyper, tefas, pitas}@aiia.csd.auth.gr

***Abstract*—This paper presents a novel facial expression recognition methodology. In order to classify the expression of a test face to one of seven pre-determined facial expression classes, multiple two-class classification tasks are carried out. For each such task, a unique set of features is identified that is enhanced, in terms of its ability to help produce a proper separation between the two specific classes. The selection of these sets of features is accomplished by making use of a class separability measure that is utilized in an iterative process. Fisher's linear discriminant is employed in order to produce the separation between each pair of classes and train each two-class classifier. In order to combine the classification results from all two-class classifiers, the 'voting' classifier-decision fusion process is employed. The standard JAFFE database is utilized in order to evaluate the performance of this algorithm. Experimental results show that the proposed methodology provides a good solution to the facial expression recognition problem.**

## I. INTRODUCTION

In recent years, developing facial expression recognition (FER) technology has received great attention [1, 2]. For the face recognition problem, the true match to the expression of a test face, out of a number of *C* different pre-determined facial expressions, is sought. This type of non-verbal communication is useful when developing automatic and, in some cases, real-time human centered interfaces, where the face plays a crucial role [3]. Examples of applications that use FER are facial expression cloning in virtual reality applications, video-conferencing, and user profiling, indexing, and retrieval from image and video databases. Facial expressions play a very important role in human face-to-face interpersonal interaction [4]. In fact, facial expressions represent a direct and naturally preeminent means of communicating emotions [5].

Recently, various methods have attempted to solve the FER problem. In [6], two hybrid FER systems are proposed that employ the 'one-against-all' classification strategy. The first system decomposes the facial images into linear combinations of several basis images using Independent Component Analysis (ICA). Subsequently, the corresponding coefficients of these combinations are fed into Support Vector Machines (SVMs) that carry out the classification process. The second system performs feature extraction via a set of Gabor Wavelets (GWs). The resulting features are then classified using CSC, MCC, or SVMs that employ various kernel functions. The method in [7] uses Supervised Locally Linear Embedding (SLLE) to perform feature extraction. Then, a minimum-distance classifier is used to classify the various expressions. SLLE computes low dimensional, neighborhood-preserving embeddings of high dimensional data and is used to reduce data dimension and extract features. The basic idea of LLE is the global minimization of the reconstruction error of the set of all local neighbors in the data set. This technique expects the construction of a local embedding from a fixed number of nearest neighbors to be more appropriate than from a fixed subspace. The Supervised-LLE algorithm uses class label information when computing neighbors to improve the performance of classification. The work in [8] introduces the ICA-FX feature extraction method that is based on ICA and is supervised in the sense that it utilizes class information for multi-class classification problems. Class labels are incorporated into the structure of standard ICA by being treated as input features, in order to produce a set of class-label-relevant and a set of class-label-irrelevant features. The learning rule being used applies the stochastic gradient method to maximize the likelihood of the observed data. Then, only class-label-relevant features are retained, thus reducing the feature space dimension in line with the principle of parsimony. This improves the generalization ability of the nearest-neighbor classifier that is used to perform FER.

This paper presents a novel FER methodology which attempts to classify any random facial image to one of the following $C = 7$ basic [9] facial expression classes: happiness (*E1*), sadness (*E2*), anger (*E3*), fear (*E4*), surprise (*E5*), disgust (*E6*), and the neutral state (*E7*). To do so, proper and unique sets of features are identified for each pair of classes. The selected features are concatenated to produce the Enhanced Feature Vectors (EFVs). This is done individually

*MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.*
*978-1-4244-4464-9/09/$25.00 ©2009 IEEE.*

for all $\frac{C(C-1)}{2}$ pair-wise comparisons between the $C$ facial expression classes. Initially, features are extracted by convolving the facial images with a set of 2-D Gabor filters of different scales and orientations. Then, a class separability measure is utilized in order select the proper subset of features for each distinct pair of classes. Then, two-class Linear Discriminant Analysis (LDA) is applied to the EFVs in order to produce the Discriminant Hyper-planes (DHs), which are essentially projections onto which large class separations are attained. The DHs are used to train the $\frac{C(C-1)}{2}$ two-class classifiers, and the corresponding two-class separations are measured. Next, the 'voting' [10] classifier-decision fusion process is employed to produce the final classification decision. This completes the proposed EFV-Classification (EFV-C) FER framework.

## II. Producing Enhanced Feature Vectors

This section presents the feature extraction process and the iterative process that is utilized in order to produce the subsets of enhanced features that compose the EFVs.

### *A. Gabor-based Feature Extraction*

Initially, a feature set that contains $M$ features is extracted from each training facial image. These $M$ features correspond to the image being convolved with $M$ 2-D Gabor filters of different scales and orientations. A 2-D Gabor filter is produced by modulating a complex exponential by a Gaussian envelope, and can allow the direction of oscillation to any angle in the 2-D cartesian plane. Thus, a filter is produced with local support that is used to determine the image's oscillatory component in a particular direction at a particular frequency. This is particularly useful for FER since different facial expressions (e.g. happiness vs. disgust, or neutral) produce these components at different directions and/or frequencies. A complex-valued 2-D Gabor function can be defined as [10]:

$$\Psi(k,x)=\frac{k^2}{\sigma^2}\exp\left(-\frac{k^2x^2}{2\sigma^2}\right)\left[\exp(jkx)-\exp(-\frac{\sigma^2}{2})\right] \quad (1)$$

To produce $M=M_s.M_o$ different Gabor functions, let us assume that $M_s$ different scales and $M_o$ different orientations are investigated. The different scales can be obtained by setting $k_i=\pi/2^i$, where $i=1,\ldots,M_s$. The different angular orientations can be obtained by selecting $M_o$ angles between 0 and 180 degrees.

### *B. Class Separability Measure for Feature Selection*

Next, a combination of the $N$ most useful features, out of the $M$ total, is selected when the task at hand is to discriminate between a specific pair of facial expression classes. Since different facial expressions produce more oscillatory components at particular directions and frequencies, it is expected that, for a given pair of classes, certain Gabor features can produce a larger class separation, than the rest of the features can. In total, there exist $\frac{C(C-1)}{2}=21$ distinct pair-wise class combinations for the $C=7$ facial expression classes: *E1-E2*, *E1-E3*, *E1-E4*, *E1-E5*, *E1-E6*, *E1-E7*, *E2-E3*, *E2-E4*, *E2-E5*, *E2-E6*, *E2-E7*, *E3-E4*, *E3-E5*, *E3-E6*, *E3-E7*, *E4-E5*, *E5-E6*, *E4-E7*, *E5-E6*, *E5-E7*, and, *E6-E7*. Thus, the feature selection process presented next creates 21 sets of enhanced features. First, the $M$ features are converted to $M$ 1-D vectors via row-concatenation, $\mathbf{f}^i$, $i=1,\ldots,M$.

Let us assume that we need to classify between a specific pair $E_x$ and $E_y$. In order to select the subset of $N$ most useful feature vectors, where $N<M$, a class separability measure that is based on the maximum value of Fisher's criterion is employed. For our purposes, this is a suitable measure since the discriminant hyper-planes that we later produce to train each two-class classifier stem from Fisher's criterion. When examining the $i-$th feature vector, this separability measure is defined as:

$$J^{\max}_{E_{x,y}}(i)=J(\mathbf{w}_{0,E_{x,y},i})=\frac{\left(\mu_{0,E_x,i}-\mu_{0,E_y,i}\right)^2}{\sigma^2_{0,E_x,i}+\sigma^2_{0,E_y,i}}, \quad (2)$$

where $\mu_{0,E_x,i}$ and $\mu_{0,E_y,i}$ denote the sample mean and $\sigma^2_{0,E_x,i}$ and $\sigma^2_{0,E_y,i}$ the sample variance of the training feature vectors of classes $E_x$ and $E_y$, respectively, when projected to the subspace defined by $\mathbf{w}_{0,E_{x,y},i}$. The discriminant vector $\mathbf{w}_{0,E_{x,y},i}$ is given by [11]

$$\mathbf{w}_{0,E_{x,y},i}=\mathbf{S}^{-1}_{W,E_{x,y},i}\left(\mathbf{m}^i_{E_x}-\mathbf{m}^i_{E_y}\right), \quad (3)$$

where $\mathbf{m}^i_{E_x}$ and $\mathbf{m}^i_{E_y}$ denote the sample mean of the feature vectors of classes $E_x$ and $E_y$, respectively, for the $i-$th feature. Moreover, $\mathbf{S}_{W,E_{x,y},i}$ is the within-class scatter matrix for the $i-$th feature, and is defined as

$$\mathbf{S}_{W,E_{x,y},i}=\sum_{\mathbf{f}^i_j\in E_x}\left(\mathbf{f}^i_j-\mathbf{m}^i_{E_x}\right)\left(\mathbf{f}^i_j-\mathbf{m}^i_{E_x}\right)^{\mathrm{T}}+\sum_{\mathbf{f}^i_j\in E_y}\left(\mathbf{f}^i_j-\mathbf{m}^i_{E_y}\right)\left(\mathbf{f}^i_j-\mathbf{m}^i_{E_y}\right)^{\mathrm{T}}. \quad (4)$$

where $j$ indicates the class (either $E_x$ or $E_y$) to which the $i-$th feature vector, $\mathbf{f}^i_j$, belongs to. So, each summation term adds up all $i-$th feature vectors that belong to a specific class.

Using (2), we now have a class separability measure that indicates how useful each of the $M$ features is. However, it is not sufficient to select the $N$ best features as the ones that produce the $N$ largest values for this separability measure. This is because each EFV, which is comprised by the $N$ features, is subsequently processed by two-class LDA to produce the discriminant hyper-plane. The concatenation of

the $N$ selected feature vectors to produce one large column vector, the EFV, is as such:

$$\mathbf{f}_j^{EFV:E_{x,y}} = \left[\mathbf{f}_j^{\mathbf{i}(1)\,\mathrm{T}}, \ldots, \mathbf{f}_j^{\mathbf{i}(N)\,\mathrm{T}}\right]^{\mathrm{T}}, \quad (5)$$

where $\mathbf{i} \underset{N}{\subset} 1, \ldots, M$, and $\mathbf{f}_j \in \mathcal{E}_x$, or, $\mathbf{f}_j \in \mathcal{E}_y$.

As a result, notions such as linear dependency between the feature vectors should be taken into account when selecting the $N$ best features. For example, if two feature vectors are linearly dependent, or close to being linearly dependent, then the selection of both these vectors, rather than only one of them, would not provide any additional benefit to the discriminant ability of the hyper-plane being produced to train the two-class classifier. For this reason, an iterative feature selection process that is again based on the class separability measure (2) is developed in order to define the group of feature vectors that should compose each pair of EFVs, for all two-class problems. Specifically, the separability measure is not applied independently to each feature but, rather, to groups of features, in order to identify the feature combination that produces the largest class separation.

### *C. Creating EFVs to Produce DHs*

The following feature selection methodology is applied in order to construct the $\frac{C(C-1)}{2}$ DHs, where each hyper-plane is associated with two realizations (one per class) of *N*-selected features. The first feature vector to be selected is the one that produces the maximum $J_{E_{x,y}}^{\max}$ value, out of all the original $M$ feature vectors, when attempting to discriminate between the two facial expression classes $E_x$ and $E_y$. Subsequently, each feature vector to be selected next is identified by creating groups of features in vector form, i.e. $\mathbf{f}_{group}$, where each group contains the feature vectors that were previously selected and a new candidate feature vector. A candidate feature vector is simply a feature vector that has not yet been selected as being one of the *N* vectors that compose the EFV. In general, if this is the $i-$th feature vector to be selected, then $M-i+1$ distinct groups of features are created:

$$\mathbf{f}_{group_i}^{j} = \left[\mathbf{f}_{selected}^{1,\ldots,i-1\,\mathrm{T}}, \mathbf{f}_{candidate}^{j\,\mathrm{T}}\right]^{\mathrm{T}}, \quad j = 1, \ldots, M-i+1. \quad (6)$$

Next, for each group of features in (6), the corresponding FLD hyper-plane that is used to discriminate between the facial expression classes $E_x$ and $E_y$ is produced:

$$\mathbf{w}_{0,E_{x,y},group_i^j} = \mathbf{S}_{W,E_{x,y},group_i^j}^{-1}\left(\mathbf{m}_{E_x,group_i^j} - \mathbf{m}_{E_y,group_i^j}\right), \quad (7)$$

where $group_i^j$ indicates that this expression only uses the group of features that are currently under consideration. Then the value of the corresponding separability measure for this group of features is calculated via (2). The selected feature is set to be the one whose corresponding group produces the maximum value of this separability measure, i.e. $J_{E_{x,y},group_i}^{\max}$.

To select all $N$ feature vectors, this process is iterated $N$ times and at its completion the $N$ selected feature vectors are concatenated to form the EFV of each class, as (5) indicates. The two EFVs that correspond to classes $E_x$ and $E_y$ are also related to a specific DH, via (7). By using this feature selection process, the two-class LDA algorithm can potentially evade problems relating to non-linear class separability. This is because multiple combinations of groups of features are examined and the group that produces the largest class separation $J_{E_{x,y},group_i}^{\max}$ is selected. Since the separability value is based on Fisher's criterion, it is expected that a combination of features that can only be used to form a strongly non-linear separation between the classes would produce a small $J_{E_{x,y},group_i^j}$ value, thus, this combination of features would be rejected. For the same reason, each EFV that is produced should not contain features that are, or are close to being, linearly dependent.

## III. Integrating Classification Results

Let us assume that a two-class classifier needs to produce a decision on whether the expression of a test image $r$ should be assigned to either the facial expression class $E_x$ or $E_y$. To do so, the test image and the two class means are projected onto the discriminant hyper-plane, $\mathbf{w}_{0,E_{x,y}}$. Then, the $L_2$ norm can be utilized to calculate the distance between the projected $r$ and the two projected class means. Subsequently, $r$ is assigned to the class associated with the smallest of the two distances.

To produce the final classification decision, i.e. determine which of the $C$ facial expression classes $r$ belongs to, results from all $\frac{C(C-1)}{2}$ two-class classifiers need to be integrated. To do so, the widely used voting classification scheme can be utilized, where the winning class for each two-class problem receives a vote and the class that accumulates the most votes is set to be the best match to the expression of the test face [10]. In case of a tie, the result that is associated with the minimum mean distance is selected.

## IV. Experimental Results

In this section, the performance of the proposed EFV-C method is evaluated and compared against contemporary state-of-the-art FER methods. The JAFFE [10] facial expression database, which contains images captured at disjoint temporal instances, has been extensively used when evaluating the classification performance of spatial facial expression algorithms. Hence, our method, as well as the spatial FER methods of [6, 7, and 8] that it is compared against, is evaluated on the JAFFE database.

A simple preprocessing step is applied to the JAFFE images before performing FER. Each face is manually cropped by taking as reference the hairline, the left and right

cheek and the chin of each face. Next, the average ratio between the vertical and horizontal dimensions of all the cropped images was calculated to be 1.28 and used to resize/down-sample the cropped images to $50 \times 39$ pixels using bicubic interpolation.

To experimentally evaluate the proposed method, we set $M_o = 6$ and $M_s = 4$, which results to obtaining $M = 24$ 2-D Gabor features that correspond to 6 different orientations and 4 different scales. Furthermore, for the enhanced-feature selection process, we set $N = 4$ heuristically (this value produced the smallest error), so each EFV is comprised by 4-selected Gabor features, out of the 24 total that are extracted.

The testing protocol that is used to evaluate the FER algorithms is the common 'leave-one-sample-out' evaluation strategy [6, 7, and 8]. During each run of this strategy, one specific image is selected as the test data, whereas the remaining images are used to train the classification system. The strategy makes maximal use of the available data for training. This process is repeated 213 times so that all the images in the database will represent the test set once. Then, the 213 classification results are averaged and the final FER rate is produced. The FER rate of the EFV-C algorithm is calculated to be 95.11%. It is noted that if all 24 features are retained, i.e. when the feature selection process that is described in section II is not applied, then this rate drops by nearly 7%. This shows that the enhanced feature vectors that are produced by the proposed feature selection process are indeed more useful when producing the facial expression classification decision. Table I summarizes the results for all competing methods and shows that EFV-C competes well with state-of-the-art solutions.

TABLE I
FER PERFORMANCE OF VARIOUS METHODS

| Method | Leave-one-sample-out FER rate |
|---|---|
| GWs+SVMs [6] | 90.34% |
| SLLE [7] | 92.90% |
| ICA-FX [8] | 94.97% |
| EFV-C | **95.11%** |

## V. CONCLUSION

A FER methodology that produces facial expression pair-specific features is proposed and its performance is evaluated. Sets of enhanced features are selected by applying an iterative process that utilizes a class separability measure. These enhanced features are then processed by corresponding two-class discriminant analysis processes in order to train all two-class classifiers. The EFV-C methodology was tested on the well-established JAFFE database under the common leave-one-out evaluation strategy. Results indicate that it provides a good solution to the FER problem by producing classification rates of 95.11%, and that it compares well with state-of-the-art methods. It is anticipated that the performance of other FER methods can be enhanced by utilizing processes that stem from this framework in order to produce high-quality features.

## ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211471 (i3DPost).

## REFERENCES

[1] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey", *Pattern Recognition*, Elsevier, vol. 36, no. 1, pp. 259-275, Jan. 2003.

[2] M. Pantic and J. M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp.1424-1445, Dec. 2000.

[3] I. S. Pandzic and R. Forchheimer, Eds., *MPEG-4 Facial Animation*. New York: Wiley, 2002.

[4] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences", *IEEE Trans. on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 36, no.2, pp. 433-449, April 2006.

[5] D. Keltner and P. Ekman, "Facial expression of emotion", in *Handbook of Emotions*, M. Lewis and J. M. Haviland-Jones, Eds. New York: Guildford, pp. 236-249, 2000.

[6] Buciu, C. Kotropoulos, and I. Pitas, "ICA and Gabor representation for facial expression recognition", in *Proc. IEEE Int. Conf. on Image Processing*, vol. 2, no. 3, pp. 855-858, Barcelona, Spain, Sep. 14-17, 2003.

[7] Liang, J. Yang, Z. Zheng and Y. Chang, "A facial expression recognition system based on supervised locally linear embedding", *Elsevier, Pattern Recognition Letters*, vol. 26, no. 15, pp. 2374-2389, Nov. 2005.

[8] N. Kwak, "Feature extraction for classification problems and its application to face recognition", *Elsevier, Pattern Recognition*, vol. 41, no. 5, pp. 1718-1734, May 2008.

[9] Ekman and W. V. Friesen, *Emotion in the Human Face*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[10] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357-1362, Dec. 1999.

[11] M. Kyperountas, A. Tefas, and I. Pitas, "Weighted piecewise LDA for solving the small sample size problem in face verification", *IEEE Trans. on Neural Networks*, vol. 18, no. 2, pp. 506-519, March 2007.