A NEW FACIAL EXPRESSION RECOGNITION TECHNIQUE USING 2-D DCT AND K-MEANS ALGORITHM

Liying Ma[†], *Yegui Xiao*[††], *K. Khorasani*[†], *Rabab Kreidieh Ward*[†††],

[†] Department of Electrical and Computer Engineering, Concordia University
1455 De Maisonneuve Blvd. West, Montreal, Quebec H3G 1M8 Canada
{liying, kash}@ece.concordia.ca
[††] Hiroshima Prefectural Women's University, Hiroshima, Japan 734-8558
[†††] Institute for Computing, Information and Cognitive Systems
University of British Columbia, Vancouver, BC, Canada

ABSTRACT

Facial expression recognition plays a vital role in realizing a highly intelligent human-machine interface, and has recently attracted much attention. In this paper, we propose a new facial expression recognition method that utilizes the 2-D DCT, k-means algorithm and vector matching. This technique is based on two main intuitive ideas: (i) complicated facial expression categories such as "anger" and "sadness", may be divided into several subcategories with different sub feature spaces where the recognition task can be performed with higher accuracy, and (ii) the k-means algorithm may be used to cluster these subcategories. A new image database with five facial expressions (neutral, smile, anger, sadness, surprise) of 60 women was constructed using a computationally efficient projection-based technique. Experimental results using the new database and an existing one (60 men) reveal that the new technique outperforms the standard vector matching technique and two recently developed methods using fixed-size and constructive one-hidden-layer neural networks. The mean recognition rate can be as high as 95% for the two databases.

1. INTRODUCTION

Recently, much attention has been paid to the computer-based recognition of facial expressions. The ultimate goal of facial expression recognition (FER) has been the realization of intelligent and transparent communications between humans and machines. So far, several FER methods have been proposed. See for examples, [1]-[4] and the references therein.

In the well-known facial action coding system (FACS) designated by Ekman [1] for facial expression description, the face is divided into 44 action units (AUs), such as nose, mouth, eyes, etc. The movement of muscles of these feature-bearing AUs are used to describe any human facial expression. The drawback of this method is that it requires 3-dimensional measurements and may thus be too complex for real-time processing. To alleviate this problem, a modified FACS using only 17 important AUs was proposed in [2] for facial expression analysis and synthesis. However, 3-dimensional measurements are still needed. The complexity of the above modified FACS is reduced when compared with the original FACS, but some information useful for facial expression recognition may be lost. In recent years, facial expression recognition based on 2-dimensional digital images has been one of the focuses of research. In paper [3], a radial basis function (RBF) neural network (NN) is proposed to recognize human facial expressions. The 2-dimensional discrete cosine transform (2-D DCT) is used to compress the entire face image and the resulting lower-frequency 2-D DCT coefficients are used to train a one-hidden-layer NN in [4].

Neural network-based recognition methods have been found particularly promising [3, 4], since NNs can easily implement complex mapping from the feature space of face images to the facial expression space. However, finding a proper network size has always been a frustrating and discouraging experience for NN developers. This is dealt with by a long and costly trial-and-error recursions. Motivated by these limitations and drawbacks, we have recently proposed to use constructive NNs [5, 6], whose recognition rates of 100% and 93.75% (without rejection), for the training and testing images, respectively have been obtained. Constructive NNs are capable of systematically determining the proper network size required by the complexity of the given problem, while reducing considerably the computational cost involved in network training when compared with the standard BP-based training techniques [4]-[7].

In all the above 2-D image based FER methods, the confusion matrices reveal that (i) expressions "smile" (smi) and "surprise" (sur) are relatively easier to recognize and are seldom confused with other expressions, and (ii) "anger" (ang) and "sadness" (sad) are very often confused, which lowers the overall recognition rate.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

Obviously, to improve the recognition performance, one needs to take measures to mitigate the confusion between "ang" and "sad". Through detailed investigations of the characteristics of each expression in the frequency domain, we have discovered that "ang" and "sad" seem to have more than one distinct subgroups that may need to be treated as individual subcategories. It is this insight that has motivated us to propose the new technique which uses the cost-efficient k-means algorithm to identify the subcategories of each expression and the vector matching scheme to perform the recognition task. Experiments using two facial image databases demonstrate that the proposed technique outperforms all the above-mentioned recognition methods while yielding attractive computational efficiency.

2. THE PROPOSED RECOGNITION TECHNIQUE

Before introducing the new technique, we show a set of sample facial expression images (size: 128×128). These front images have 5 expression categories with four of them ("smile", "anger", "sadness", and "surprise") being the subject of the recognition problem. More details will be given in the next section regarding the databases used in the experiments.

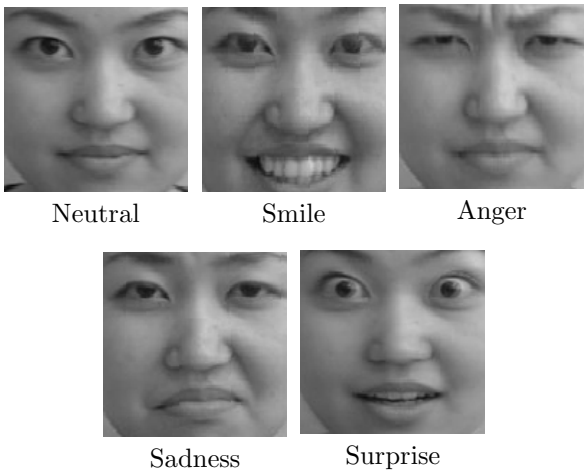

Fig. 1 Sample images from the database used.

The proposed technique is divided into two phases: training and testing, which are described separately below.

(A) Training

The training process consists of the following elements:

E1 Difference images

The features of facial images used in recognition must not be influenced by the appearance of any individual human. Therefore, preprocessing of the face images is needed in order to extract some information that is required by the recognition task and shared by all the expression images of the same category. One may make difference images by subtracting the neutral images from the expression images. The difference images are then expected to have much less to do with the appearance of the human whose facial expressions are the subject of recognition.

E2 Data compression using 2-D DCT

Obviously, it is very difficult for the classifier to recognize the facial expression from the difference images, as a difference image still has a large number of data. To facilitate the recognition, we need to compress the difference image for reducing the number of data in a proper way, without losing the key features that play important role in the recognition task. The 2-D DCT used frequently in image compression is a powerful tool for this purpose. The 2-D DCT can reduce the number of data significantly by transforming an image into the frequency domain where the lower frequencies present relatively large magnitudes while the higher frequencies indicate much smaller magnitudes. That is to say, the higher frequency components can be ignored without damaging the key characteristics of the original difference image, as far as the facial expression recognition is concerned. The 2-D DCT coefficients of a square with size $L_1 \times L_2$ of the lower frequencies (Fig.2) hold much of the information on the facial expressions, and are arranged as a vector for training or testing purposes.

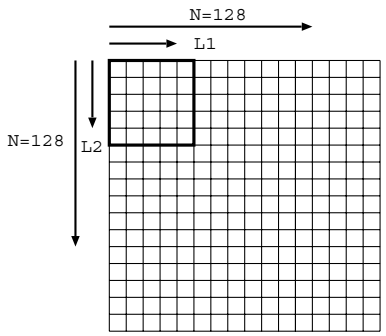

Fig. 2 2-D DCT coefficients of lower frequencies.

E3 Subgroup search using k-means algorithm

As mentioned before, the expressions "ang" and "sad" have many distinct subgroups (clusters), and the recognition rates for these two expressions may be improved if one treats them individually, instead of combining them into a single (global) group. To this end, one needs to use a tool to specify the members of each subgroup. In this work, the basic k-means algorithm is used for this purpose. In each search, the initial vectors (centroids) for each subgroup are randomly selected from the training vectors, and the nearest-neighbor (minimum distance) rule is utilized in classifying the group members. For a user-specified number of subgroups, a number of independent search runs are performed, and

the outcome of a run with minimum summed distance is taken as the search result. The number of independent search runs is set large enough such that the outcome remains unchanged. Because the dimension of each training vector is small, the k-means algorithm converges very fast. The centroids for all the subgroups are regarded as the standard vectors for recognition. As will be seen from the experimental results in the next section, a single group seems enough for expressions "smi" and "sur".

E4 Two-step recognition process

Once the standard vector(s) for each expression are obtained, the recognition is performed based on vector matching. The expression represented by the standard vector which indicates minimum distance to the vector being recognized is regarded as the outcome of recognition. Concretely, the recognition of an expression image is divided into the following two unique steps in order to reduce the computational cost.

Step 1: The vector to be recognized is handled in a feature space where each of the 4 expression categories has only one standard vector. If the vector is classified as "smi" or "sur", then the recognition is achieved. If this is not the case, then go to step 2.

Step 2: The vector is handled in another feature space where there are only two expression categories: "ang" and "sad", and there are multiple standard vectors corresponding to the subgroups of each category. The classification result is regarded as the final outcome of the recognition.

(B) Testing

After the training is completed, a different portion of the facial expression images is used to test the performance of the trained standard vectors.

3. EXPERIMENTAL RESULTS

3.1. A new database

In this work, a new database is constructed using an efficient projection-based procedure. The database consists of images of 60 women, with each having 5 expression images, i.e., neutral, smile, anger, sadness, and surprise. This is similar to the database we used in our previous work [4, 5, 6].

A digital camera is used to take frontal images of each person. The images are incorporated into the computer where they are converted into gray images of size 168×168. An image is first divided into top half and bottom half blocks with equal size (84×168). The top block is further divided into two sub-blocks from the vertical middle, each having a size 84×84. Then, horizontal and vertical projections (i.e., summations of the gray-level values of the pixels on the same horizontal or vertical line) for the top two sub-blocks are performed. The minimum points of the projection curves will be the candidates for the eye positions. To get stable results, the DFT is used to smooth the curves (only 8 DFT coefficients are used in the IDFT). A set of projection curves is given in Fig. 3, as an example. Clearly, the eye positions are correctly detected and determined. Next, the mouth is detected using similar projections applied to the bottom block. To obtain reliable mouth positions, compensation of white teeth is introduced before the projections are performed, by setting a proper empirical threshold such that the white teeth are detected and blackened. Based on the eye and mouth positions detected, the image is rotated and scaled if needed, and finally an image of size 128×128 is produced. The image samples shown in Fig.1 belong to this new database.

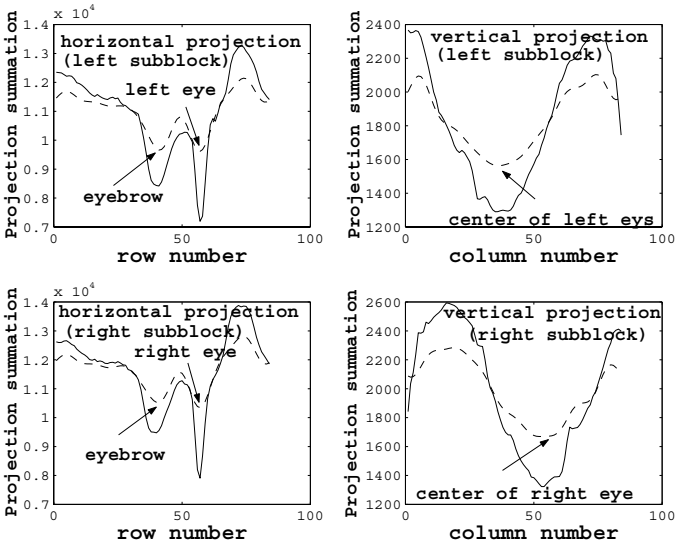

Fig. 3 Projection curves for eye position detection, solid line: original, broken line: smoothed.

3.2. Experimental results

First, the testing recognition rates of the 4 expressions corresponding to the new database are shown in Fig. 4 with respect to the number of groups of "ang" and "sad", for both training (40 samples) and testing (20 samples). In all the experimental results shown in this work, the block size is set to $L_1 = L_2 = 8$, which has been found suitable for the k-means algorithm to work effectively. It is obvious that the recognition rates for "ang" and "sad" increase at the beginning as the number of groups increases, and then become nearly saturated when the number of groups are above some threshold. If the number of groups for "ang" and "sad" is set to 10, the mean testing recognition rate is 95%, while the corresponding result using only a single group is 85%. The corresponding confusion matrix for testing is now provided in Table 1. Clearly, using more groups can mitigate the confusion between "ang" and "sad".

Tab 1 Confusion matrix of testing with 10 groups.

exp	smi	ang	sad	sur
smi	20	0	0	0
ang	0	**19**	0	1
sad	0	0	**18**	2
sur	1	0	0	19

Tab 2 Comparison of mean recognition rates ((a): men's database, (b):women's database).

method	training rate (%)	testing rate (%)
vector matching (a)	86.9	85.0
fixed-size NN [4] (a)	100.0	89.7
constructive NN [5] (a)	99.4	93.8
proposed technique (a)	94.4	**93.8**
vector matching (b)	86.9	85.0
proposed technique (b)	93.8	**95.0**
proposed technique (a+b)	95.9	**95.6**

Similar results are obtained for the men's database as described in [4, 5, 6].

In the next set of simulations we select 80 samples for training which consisted of 40 samples from men's database and 40 samples from women's database. The remaining 40 samples of the two databases are used as testing samples. The mean recognition rates for training and testing are plotted in Fig. 5 with respect to the number of groups of "ang" and "sad". It can be seen from Fig. 5 that the recognition rates improve as the number of groups increases. When the number of groups is set to 25, the best testing mean recognition rate of 95.6% is achieved, which is the best result that we have ever obtained.

To summarize, the testing recognition rates obtained by our proposed technique have been improved as compared to all the previous results published in the literature using simple vector matching, fixed-size NN, and the constructive one-hidden-layer NN [4, 5, 6]. Comparisons of mean recognition rates for the previous methods and the proposed technique is given in Table 2.

4. CONCLUSIONS

In this work, a new FER technique is proposed which is based on two intuitive ideas obtained from extensive analyses of facial expression characteristics. In addition to being computationally efficient, the new technique that utilizes the 2-D DCT and the k-means algorithm works surprisingly well. Extensive experiments have been conducted to demonstrate the superior effectiveness of the new method. Combining our new technique with a proper NN structure is a topic for future research.

5. REFERENCES

[1] P. Ekman and W. Friesen, Facial Action Coding System, Consulting Psychologists Press, 1977.

[2] F. Kawakami, H. Yamada, S. Morishima and H. Harashima, "Construction and Psychological Evaluation of 3-D Emotion Space," Biomedical Fuzzy and Human Sciences, Vol.1, No.1, pp.33–42, 1995.

[3] M. Rosenblum, Y. Yacoob, and L. S. Davis, "Human expression recognition from motion using a radial basis function network architecture," IEEE Trans. on Neural Networks, Vol.7, No.5, pp.1121-1138(Sept. 1996).

[4] Y. Xiao, N. P. Chandrasiri, Y. Tadokoro, and M. Oda, "Recognition of facial expressions using 2-D DCT and neural network," Electronics and Communications in Japan, Part 3, Vo.82, No.7, pp.1-11(July, 1999).

[5] L. Ma, K. Khorasani, *"Facial expression recognition using constructive neural networks,"* Proc. SPIE, vol.4380, pp.521-530 (2001),

[6] L. Ma, K. Khorasani, *"Facial expression recognition using constructive feedforward neural networks,"* IEEE Trans. System, Man, and Cybernetics, in press.

[7] T. Y. Kwok and D. Y. Yeung, *"Objective functions for training new hidden units in constructive neural networks,"* IEEE Trans. on Neural Networks, Vol.8, No.5, pp.1131-1148(1997).

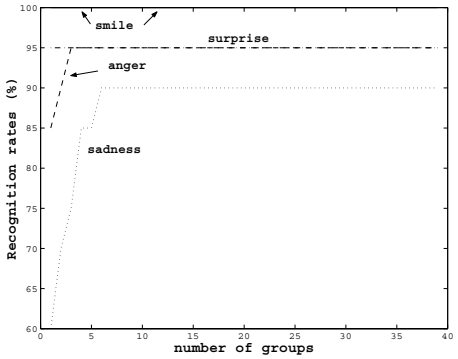

Fig. 4 Testing recognition rates for the new database as a function of the number of groups.

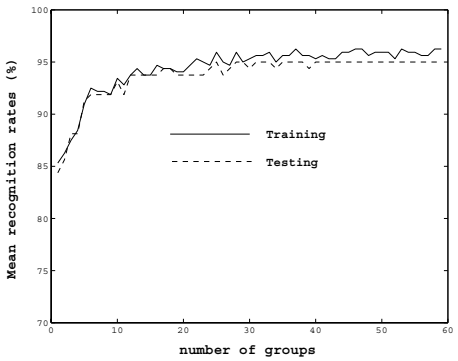

Fig. 5 Mean recognition rates for the two databases as a function of the number of groups.