Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Facial Expression Recognition Using Embedded Hidden Markov Model

Languang He, Xuan Wang, Member, IEEE, Chenglong Yu, Member, IEEE, Kun Wu
Intelligence Computing Research Center
HIT Shenzhen Graduate School
Shenzhen, China
{telent, wangxuan, ycl, wukun} @cs.hitsz.edu.cn

***Abstract*—Embedded Hidden Markov Model (EHMM) has been applied to many areas due to its excellent features. In this paper, we present a novel method for Facial expression recognition by using the EHMM. We use five scales and eight orientations Gabor features to represent the expression image. Further, we use the EHMM to recognize the facial expression. In the EHMM structure, the super states are used to model the expression image along vertical direction while the inner states are used to model the expression image along horizontal direction. Our test results and analysis based on the JAFFE database demonstrate that the proposed method is effective and achieves higher average recognition accuracy (96.16%).**

***Keywords*—Facial expression recognition, Gabor wavelet, Embedded Hidden Markov Model**

I. INTRODUCTION

A growing interest in human-computer interaction systems has been developed in the last decade. As an important part of intelligent and interactive system, Facial expression recognition (FER) has become more important. Since Ekman and Friesen presented that the human emotions can be divided into six primary emotions (happiness, anger, fear, surprise, sadness, disgust), a great amount of methods for the FER had been proposed. All the methods of FER can be divided into two types, one is based on the still image and the other is based on the dynamic video. For the former type, the information of recognizing the facial expression is just from the single still image and the available information is relatively limited. The important methods in literature include the Eigenface method [1], Independent Component Analysis (ICA) [2], Linear Discriminant Analysis (LDA) [3], Local Binary Pattern (LBP) [4], Wavelet transformation [5] and others. For the second type of methods, more useful information can be obtained by dealing with a sequence of intensity image frames. The successful methods include Optical Flow Models [6] and Hidden Markov Models [7] [8], which achieved good recognition results. According to the existing practical applications, the former type is relatively simple and can satisfy the real-time requirement, the latter type costs high and cannot satisfy real-time requirement.

Most FER methods extract expression features from the whole image or from parts of image like forehead, eyes, nose, mouth, and then combine them together. Many features used by these methods can represent the image well, but most of the classification methods don't efficiently model the position relationships of the image's parts. In fact, the separate parts of the image occur in the particular order, such as: eyes are up to mouth, mouth is up to the chin, and left eye is left to right eye. In this paper, we present a novel method for FER. We use Gabor features to represent the face image and use EHMM to recognize the facial expression. The method can model the image along vertical and horizontal directions more efficiently.

In our method, first we preprocess the face images and resize them to 30 x 35 in pixels, then we extract five scales, eight orientations Gabor features of the image as the image representation. Second we use the Embedded Hidden Markov Model (EHMM) to recognize. Each EHMM includes seven super states, along with 6 or 9 inner states for each inner HMM. Moreover, the super states are used to model the image from top to down, and the inner states to model the image from left to right.

The rest of this paper is organized as follows: Section Ⅱ introduces the Gabor wavelet theory and feature extraction method. In Section Ⅲ, we introduce EHMM and present a novel FER method by using the EHMM and Gabor transformation. In Section Ⅳ, we provide the experimental results and analyze the proposed method. Finally, concluding comments are presented in Section V.

II. GABOR FEATURE EXTRACTION

A. Gabor wavelet

The Gabor wavelet transform [9] is one of the best schemes for image representation. Its major advantage is that it achieves the lower bound on the joint entropy. The majority of receptive field profiles of the mammalian visual system match quite well to this types of function [10]. Based on its fine property, the Gabor transformation has been used widely in image processing applications, such as texture segmentation, image compression, and face recognition.

A family of 2-D Gabor kernels [11] is the product of a Gaussian envelope and a plane wave, defined as (1):

$$\psi_{u,v}(x,y)=\frac{k_{u,v}^2}{\sigma^2}\exp\left(-\frac{k_{u,v}^2\left(x^2+y^2\right)}{2\sigma^2}\right)\bullet\left[\exp\left(ik_{u,v}\bullet\begin{pmatrix}x\\y\end{pmatrix}\right)-\exp\left(-\frac{\sigma^2}{2}\right)\right] \quad (1)$$

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

Here, (x, y) is the variable in spatial domain and $k_{u,v}$ is the frequency vector, which determines the scales and the orientations of Gabor kernels. In the proposed system, we define $k_{u,v}$ as following:

$$k_{u,v} = \begin{pmatrix} k_v \cos\varphi_u \\ k_v \sin\varphi_u \end{pmatrix} \quad (2)$$

Where $\varphi_u = (u * \pi)/8$, $k_v = \pi \exp(-(v+2)/2)$, u and v are orientation factor and scale factor respectively, different value of subscript u and v represents different Gabor kernel.

B. Feature extraction

Based on the multiplication-convolution property, the Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier transform of the harmonic function and the Fourier transform of the Gaussian function. Give an image I(x, y), its Gabor transformation at particular position can be computed by a convolution with the Gabor kernel:

$$\mathrm{Gabor}_{u,v}(x,y) = I(x,y) * \psi_{u,v}(x,y) \quad (3)$$

The two-dimensional Gabor filter is composed of the real part and imaginary part. We can compute the real part (4) and imaginary part (5) by using (1). In the proposed system, we use the magnitude (6) to get Gabor kernel.

$$\mathrm{real}(x,y) = \frac{k_{u,v}^2}{\sigma^2} \exp\left(-\frac{k_{u,v}^2 (x^2+y^2)}{2\sigma^2}\right) \bullet \left[\cos\left(k_{u,v} \bullet \begin{pmatrix} x \\ y \end{pmatrix}\right) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \quad (4)$$

$$\mathrm{img}(x,y) = \frac{k_{u,v}^2}{\sigma^2} \exp\left(-\frac{k_{u,v}^2 (x^2+y^2)}{2\sigma^2}\right) \bullet \sin\left(k_{u,v} \bullet \begin{pmatrix} x \\ y \end{pmatrix}\right) \quad (5)$$

$$\mathrm{mag}(x,y) = \sqrt[2]{\mathrm{real}^2(x,y) + \mathrm{img}^2(x,y)} \quad (6)$$

A group of Gabor kernels will be built by using different scales and orientations. The kernels are convolved with the images resulting in the Gabor space. This process is closely related to the activity in the primary visual cortex. We choose v = {0, 1, 2, 3, 4}, u = {1, 2, 3, 4, 5, 6, 7, 8} to denote the five scales and eight orientations, i.e. from $\pi/8$ to π single step of $\pi/8$. Finally, we obtain 40 Gabor kernels and use them to perform Gabor transformation. One of face images and its Gabor features are shown in Fig.1.

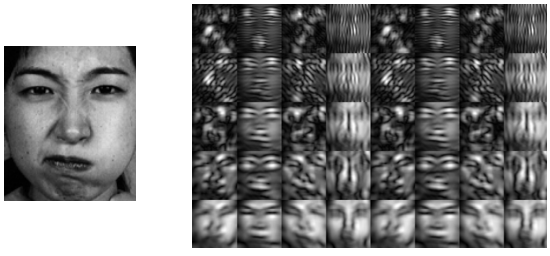

Figure 1. The face image and its Gabor features

III. EHMM RECOGNITION

A. Embedded Hidden Markov Model

A HMM [12] provides a statistical model for a set of observation data sequences. It includes two forms of stochastic finite process. First is the Markov chain of finite state, which describes the transfer from one state to another. Second one describes the probabilities between states and observation data. For statistical characterize a HMM, state transition probability matrix, an initial state probability distribution and a set of probability density functions associated with the observation data for each state are essential.

Typically HMM is a 1-D structure suitable for analyzing 1-D random signals, e.g. speech signals. A 1-D HMM can be developed into the pseudo 2-D structure [13] by extending each state in 1-D HMM as a sub HMM, which is shown in Fig.2. In this way the HMM consists of a set of super states along with the set of inner states. Such a pseudo 2-D HMM is also called Embedded HMM. The super states are used to model 2-D data along one direction with the inner states to model 2-D data along the other direction.

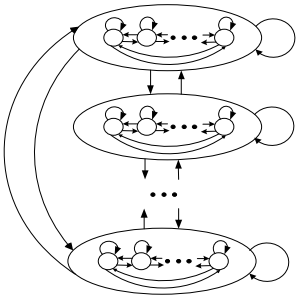

Figure 2. Topology structure of Embedded HMM

B. EHMM Structure

In the proposed method, we define the structure of EHMM as shown in the right part of Fig.3.The EHMM includes seven super states, and each inner HMM is composed of 6 or 9 inner states. The transitions in a vertical direction are only allowed between two adjacent super states which include the super state itself. The transitions in a horizontal direction are only allowed between two adjacent inner states in a super state including the inner state itself.

Based on the EHMM structure, we segment the face image like the face segmentation in Fig.3, i.e. we divide the face image into seven parts from top to bottom. Further, we cut the first part and seventh part of image into the six smaller parts. While from second part to the sixth part of face image, we cut them into the nine smaller parts. From top to bottom, the seven

parts of face image features are distributed into the seven super states. From left to right, the smaller parts of face image features are distributed into the corresponding inner states.

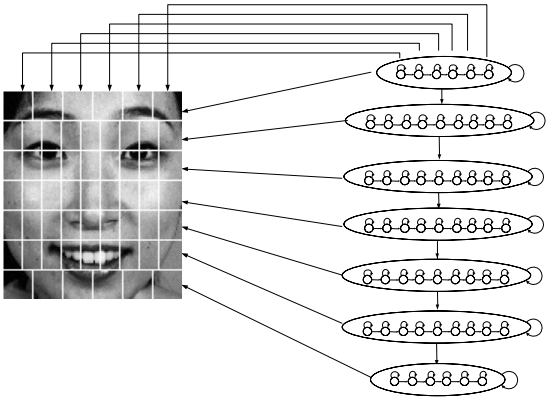

Figure 3. The EHMM structure and corresponding image segmentation

C. EHMM training and recognition

To recognize seven expressions (six basic facial expressions and one neutral), we build EHMM for each of them. In the expression database, we randomly choose two images of each expression from ten persons as our training set. In this way, for each model we can get the 20 training images and totally we get 140 training images.

For the model training, first we extract features from training set and distribute them to EHMM. After initializing the model parameters, we begin to r-estimate the model parameters by using consecutive iterations. In each iteration we compute the Viterbi likelihood of the model, if the Viterbi likelihood is smaller than a threshold or number of iterations is bigger than a number, we judge that the model parameters are converged and lead to the completion of EHMM training. This process is shown in the left part of Fig.4.

The initial parameters of EHMM are set as following:

- Number of super states: 7;
- Number of inner states: 6, 9, 9, 9, 9, 9, 6;
- Gaussian Mixture: 9;
- Number of training images: 20;
- Length of observed value: 40;
- Threshold: 0.00001;
- Max iteration number: 80;

For recognition, we extract the feature sequence O of expression image and distribute it to each expression model respectively, then we compute the Viterbi likelihood probability $p(O \mid m_i)$ for each expression model. An expression image is classified to k if and only if:

$$p(O \mid m_k) = \operatorname{argmax} p(O \mid m_i) \quad i = 1,2,3,4,5,6,7 \qquad (7)$$

This process is shown in the right part of Fig.4.

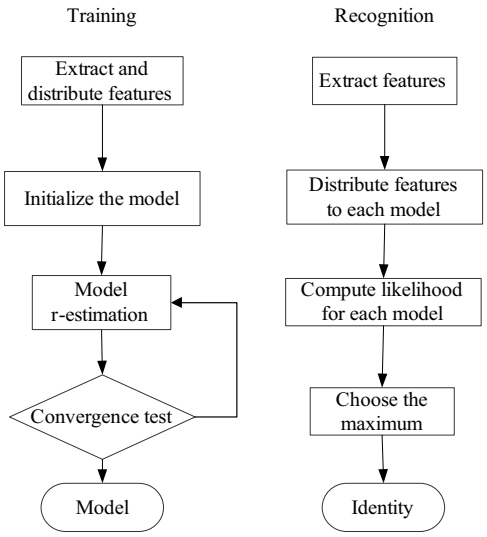

Figure 4. The scheme of EHMM training and recognition

IV. TESTS AND ANALYSIS

A. Data set and test environment

In this paper, we use the JAFFE Database to test the proposed method. The database contains 213 images of 7 facial expressions posed by 10 Japanese female models. For each person, there are 2 to 4 images for each expression. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. The database was planned and assembled by Miyuki Kamachi, Michael Lyons, and Jiro Gyoba.

The machine configuration for the tests is Pentium (R) 4 CPU 3.00GHz, 1.5GB memory.

B. Test Results

The JAFFE database contains 213 images. We choose 140 images as training set, and we have remaining 73 images to recognize in one test. In each test, we train the models and recognize the testing set. After repeating this procedure five times, we get the test results given in Tab.1. The results indicate that our method obtain a higher recognition accuracy.

TABLE I. RESULTS OF RECOGNITION

Emotion labels	Total images	Recognized images	Recognition accuracy
anger	50	50	100%
disgust	45	42	93.33%
fear	60	56	93.33%
happiness	55	54	98.18%
neutral	50	50	100%
sadness	55	51	92.72%
surprise	50	48	96.00%
Total	365	351	96.16%

C. Comparison

In order to make a comparison, we make another test by using the classical KNN method. For KNN (k=1) recognition, we use the same features as we use in the EHMM method. The

comparative results are shown in the Fig.5. We can see that EHMM method is more effective and its average recognition accuracy is ten percent higher than that of the KNN method (the average recognition accuracy of KNN is 86.02%)

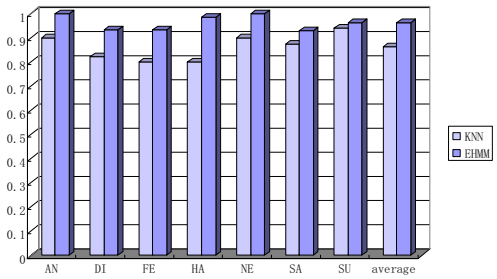

Figure 5. The comparison of KNN and EHMM

D. *Performance Analysis*

For proposed method, the face image size of feature extraction is 30 x 35 in pixels. By performing the five scales, eight orientations Gabor transformation, the size of each image's feature become 30 x 35 x 40 = 42000. The features are bigger and we make the performance analysis. We extract features of each image three times, and recognize random image 1000 times by using our models. In this way, we obtain that the average time of extracting features from one image is 34.94 ms, and the average time of recognizing one image's feature is 692.83 ms. Therefore, in order to recognize a face image, we totally need to spend about 727.77 ms. It costs a slight high and can not satisfy the real-time requirement well. There exists a tradeoff between recognition accuracy and cost thus we have to find the balanced point which has the good recognition accuracy and lower cost.

For the purpose of analysis, we use six other datasets; in first dataset, we resize the face images to 27 x 32 in pixels, and the image sizes of second to sixth datasets are given in pixels as 24 x 28, 21 x25, 18 x21, 15 x 18, and 12 x 14, respectively.

For each dataset, we train models and recognize expression by using our method. Base on our statistics, the time of feature extraction and recognition is shown in Fig.6, where Extracting time denotes the average time of extracting features from one

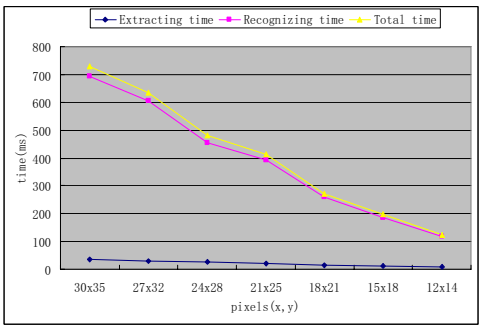

Figure 6. The time cost of different dataset

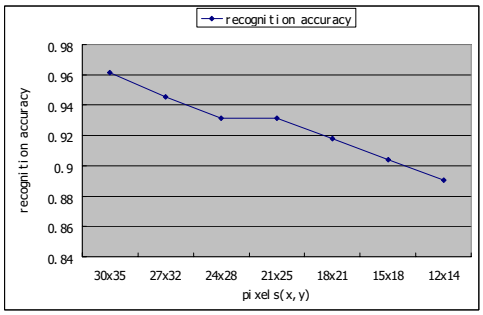

Figure 7. The recognition accuracy of different dataset

image. Whereas, Recognizing time is the average recognizing time of one image's feature, and Total time is the sum of Extracting time and Recognizing time. The recognition results are shown in Fig.7. From Fig.6 and Fig.7, it is clear that we can get better efficiency and keep recognition accuracy above 93% when image size is about 21 x 25.in pixels. And we can use it to face a variety of application environments.

V. CONCLUSION

Facial Expression Recognition has been investigated over years, many methods have been proposed and some have achieved good recognition results. In this paper, we have presented a novel method for FER based on EHMM. We made test and performance analysis for proposed method; the results and comparison demonstrated the effectiveness of proposed method. In addition, our system has two features:

1) The pretreatment is simple, which simply perform the Gabor transformation to the face image.

2) The EHMM has good efficiency for FER. Using the method, we can recognize 21000-dimensional feature in 0.41s and keep the recognition accuracy above 93%.

ACKNOWLEDGMENT

This work is supported by the National High-tech R&D Program of China (863 Program, No. 2007AA01Z194). And we thank M. Lyons for providing the JAFFE database.

REFERENCES

[1] Turk M, Pentland A. Eigenfaces for recognition [J]. Journal Cognitive Neuro - science, 1991, 3 (1) : 71～86.

[2] Liu C, Wechsler H. Independent component analysis of Gabor features for face recognition [J]. IEEE Transactions on Neural Networks, 2003, 14 (4) : 919～928.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711–720, Jul. 1997.

[4] S.Marcel, Y.Rodriguez and G.Heusch, On the Recent Use of Binary Patterns for Face Authentication, International Journal of Image and Video Processing, Swiss, 2006.

[5] Yongzhao ZHAN, Jingfu YE, Dejiao NIU, Peng CAO. Facial Expression Recognition Based on Gabor Wavelet Transformation and Elastic Templates Matching. Proceedings of the Third International Conference on Image and Graphics, 2004.

[6] Yacoob Y, Davis L. Recognizing human facial expressions form long image sequences using optical flow[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16 (6) : 636～642.

[7] Otsuka T, Ohya J. Recognizing multiple persons facial expressions using HMM based on automatic extraction of significant frames from image sequence[A]. In: Proceedings of the International Conference on Image Processing[C] , California, USA, 1997: 546～549.

[8] M. Pardas, A. Bonafonte. Facial animation parameters extraction and expression recognition using Hidden Markov Models. Signal Processing: Image Communication 17:675-688,2002.

[9] D. Gabor, "Theory of communications," J. Inst. Elec. Eng., vol. 93, pp.429-457, 1946

[10] Lee T S. Image representation using 2D Gabor wavelets [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18 (10) : 959～971.

[11] M. Lyons, S. Akamatsu, etc. Coding Facial Expressions with Gabor Wavelets. Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara Japan, 200-205, 1998

[12] Rabiner L, "A tutorial on HMM and selected applications in speech recognition", Proc. IEEE, Vol. 77, No. 2, pp. 257-286, 1989.

[13] S. Kuo, O. Agazzi, "Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov models," IEEE Trans. Pattern Anal. Machine Intell. 16 (8) (1994) 842–848