Applied Mathematics & Information Sciences
An International Journal

Multi-Modal Emotion Recognition Fusing Video and Audio

Chao Xu[1], *Pufeng Du*[2], *Zhiyong Feng*[2], *Zhaopeng Meng*[1], *Tianyi Cao*[2], *and Caichao Dong*[2]

[1] School of Computer Software, Tianjin University, 300072 Tianjin, China
[2] School of Computer Science and Technology, Tianjin University, 300072 Tianjin, China

Received: 7 Sep. 2012; Revised 15 Nov. 2012; Accepted 18 Nov. 2012
Published online: 1 Mar. 2013

Abstract: Emotion plays an important role in human communications. We construct a framework for multi-modal fusion emotion recognition. Facial expression features and speech features are respectively extracted from image sequences and speech signals. In order to locate and track facial feature points, we construct an Active Appearance Model for facial images with all kinds of expressions. Facial Animation Parameters are calculated from motions of facial feature points as expression features. We extract short-term mean energy, fundamental frequency and formant frequencies from each frame as speech features. An emotion classifier is designed to fuse facial expression and speech based on Hidden Markov Models and Multi-layer Perceptron. Experiments indicate that multi-modal fusion emotion recognition algorithm which is presented in this paper has relatively high recognition accuracy. The proposed approach has better performance and robustness than methods using only video or audio separately.

Keywords: Emotion Recognition, Multi-modal Fusion, HMM, Multi-layer Perceptron.

1. Introduction

Emotion recognition is an important research field of pattern recognition. Emotion takes a significant role in human communications, and has an effect on perception and decision making. Research on emotion recognition is related to many subjects, such as computer vision, speech signal processing, artificial intelligence, psychology, sociology, and so on. Emotion recognition is widely used in human-computer interaction, medical care, security, communication, and many other fields. Psychological researchers have found six kinds of affective states, including happiness, anger, sadness, fear, surprise, and disgust [1]. Many researches on emotion recognition have been focused on images or image sequences with facial expressions. Other researchers have recognized affective states of speech signals.

Research on social psychology indicates that facial expression is an important modality in human communications [2]. Franco L took face images in Yale database as training sets to recognize facial expressions [3]. Ying Z applied Principal Component Analysis and Linear Discriminating Analysis to reduce data dimension, and made use of Support Vector Machine to recognize affective states [4].

Experiments conducted by Bassili J show that people are more expert in identifying facial expressions in image sequences than in static images [5]. Researchers have proposed many methods to recognize affective states of image sequences [6]. Facial expressions are derived from motions of facial muscles. Ekman P and Friesen W developed Facial Action Coding System to code for muscle contraction [7]. Yu M calculated optical flows of facial image sequences to represent facial expressions, and took Artificial Neural Network as emotion classifier [8]. Yang G introduced Hessian Matrix into Lucas-Kanade Optical Flow, and used Hidden Markov Models to recognize facial expressions [9].

Many researchers have used speech signals to recognize emotions of people [10]. Zhao L put forward an emotion recognition method based on speech. He extracted energy, fundamental frequency and formant frequencies from each speech frame, and calculated their statistics, such as mean and variance [11]. Lin Y extracted 39 candidate features from speech signals, and applied Sequential Forward Selection to select 5 features as the best feature sub-

* Corresponding author: {pdu, mengzp}@tju.edu.cn

set [12]. Ye J calculated Hurst index and multiple parting spectrum parameters as speech features, and used Support Vector Machine to classify emotions [13].

Psychological studies indicate that humans identify affective states mainly based on facial expressions and speeches. As a result, emotion recognition is inherently an issue of multi-modal analysis. Three fusion strategies have been applied in multi-modal emotion recognition: feature-level fusion, model-level fusion, and decision-level fusion. Feature-level fusion concatenates speech features and facial expression features to construct combined feature vector. However, features from different modalities do not always have the same time scale. The difficulty in synchronization of video and audio streams influences the performance of feature-level fusion. Model-level fusion relaxes the requirement of synchronization and makes use of the correlation of multi-streams as well. Zeng Z presented Multi-stream Fused Hidden Markov Model (MFHMM), in which each group of features was modeled by a component HMM. MFHMM acquires a good balance between model complexity and performance [14]. Most of the existing works on multi-modal emotion recognition apply decision-level fusion, which independently models features from video and audio, and combines these unimodal recognition results in the end. Go J presented a bi-modal emotion recognition approach which calculated the weighted sum of recognition results from video and audio [15].

In this paper, we propose a multi-modal approach which fuses expression features and speech features. Emotional features are extracted from facial image sequences and speech signals, and fused on decision level. In the phase of feature extraction, an Active Appearance Model is constructed for facial images to locate and track facial feature points. Facial Animation Parameters (FAPs) are computed as expression features based on motions of facial points. Fundamental frequency, short-term mean energy and formant frequencies are calculated from each speech frame as speech features. In the phase of emotion recognition, Hidden Markov Models are constructed for every expression and emotional speech. Multi-layer Perceptron is applied to fuse expression features and emotional speech features, and affective states are acquired.

2. Affective feature extraction

Feature extraction is a critical procedure of pattern recognition issue. The accuracy of the extracted affective features has a significant impact on the result of emotion recognition. We extract facial expression features and emotional speech features respectively from video sequences and speech signals. In order to locate and track facial feature points, we construct an Active Appearance Model for facial images. Facial Animation Parameters are calculated as expression features based on motions of facial feature points. We extract pitch, energy and other features from emotional speech.

2.1. Active Appearance Model for facial expression images

Active Appearance Model was first proposed by T. F. Cootes. It is widely used in detecting and tracking feature points of flexible objects, especially facial feature points. It can also be used in image reconstruction. Active Appearance Model is an extension of Active Shape Model. In Active Shape Models, shapes are modeled for training images. In Active Appearance Models, shape and grey-level information are combined for modeling. As a result, Active Appearance Model is more accurate than Active Shape Model.

We build an Active Appearance Model for facial expression images. Images in the training set have different light intensities, postures and facial expressions. Locations of facial feature points have been previously marked on each training image. Figure 1 shows several facial images in Cohn-Kanade 2010 Facial Expression Database [16].

Firstly, we build shape models for training images. In order to compare corresponding points of different shapes, we align all shapes in the training set. Aligning procedure is implemented by scaling, rotating and translating. A least-square approach is used in aligning a pair of shapes. The mean shape of all aligned shapes is acquired.

Figure 1 Cohn-Kanade 2010 Facial Expression Database

Secondly, grey-level models are constructed for the training images. A triangulation algorithm is applied to make facial feature points in training images align to points in the mean shape. Figure 2 shows the processing result of triangulation algorithm and mean grey-level appearance.

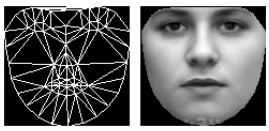

Figure 2 Active Appearance Model

For a given image, we use an iteration algorithm to acquire a reconstructed image which is similar to the original image. At the same time, we get the locations of facial feature points. In the experiments, we apply AAM-Library to implement the training process of Active Appearance Model, and facial feature point detection.

Figure 3 shows the image reconstruction result and the located facial points, where the left image in each line is the original facial image with previously marked points, the middle image is the reconstructed image, and the right image is the facial image with detected points.

Experiments show that the choice of training images has an important impact on the performance of image reconstruction and feature point detection. If the training set contains only images with neutral expressions, then the reconstructed images will always be wrong. If the training set contains only images with slight expressions, the processing result of images with intensive expressions will not be accurate. So we select images with different facial expressions of several volunteers to build up the training set. The method has a relatively good result in detecting facial feature points.

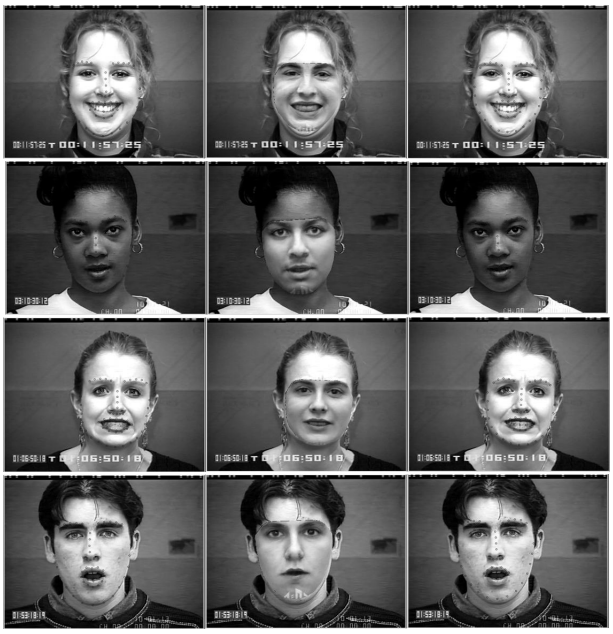

Figure 3 Facial feature point locating

2.2. Expression feature extraction based on Facial Animation Parameters

After detecting facial feature points from image sequences, we calculate Facial Animation Parameters based on motions of feature points. Facial Animation Parameters are a set of parameters compliant with ISO MPEG-4 standard. They are related with motions of facial muscles, and they are able to express basic facial motions, such as frown, blinking eyes, opening or closing mouth, and so on. So Facial Animation Parameters can represent most facial expressions. Facial Animation Parameters are calculated from horizontal or vertical displacement of facial feature points. For instance, FAP 19 is acquired from vertical displacement of top left eyelid, FAP 31 is computed from vertical displacement of left inner eye brow, and FAP 51 is acquired from vertical displacement of top middle outer lip.

In the experiments, we use Facial Animation Parameters as expression features, including FAP 19, 20, 21, 22 which are related to motions of eyelid, FAP 31, 32, 33, 34, 35, 36, 37, 38, 39 which are related to motions of brow, and FAP 51, 52, 53, 54, 55, 56, 57, 58, 59, 60 which are related to motions of mouth. Table 1 describes FAP 19, 31 and 51 in detail. Facial Animation Parameters can be standardized by Facial Animation Parameter Units (FAPUs). Facial Animation Parameter Units enable FAPs of all facial images to be evaluated in the same way. Facial Animation Parameters are able to distinguish different kinds of facial expressions, and they can be easily and reliably calculated. As a result, they can be used as expression features to recognize affective states.

Table 1 Facial Animation Parameters

No.	FAP Description	FAPU
19	Vertical displacement of top left eyelid	ENS
31	Vertical displacement of left inner eye brow	ENS
51	Vertical displacement of top middle outer lip	MNS

2.3. Emotional speech feature extraction

Speech makes a big contribution to representing emotions in human communications. We can judge peoples affective states through pitch and loudness of speech. The same content of speech can be expressed in different ways.

We extract fundamental frequency, formant frequencies and short-term mean energy from each frame of speech signals as emotional speech features.

Short-term mean energy is defined as convolution of square of speech amplitude and square of window function. The formula of short-term mean energy is as follows.

$$E_n = \sum_{m=n}^{n+(N-1)} x^2(m)\omega^2(n-m). \quad (1)$$

where N is the length of a frame, $x(\cdot)$ is the amplitude of speech signals, $\omega(\cdot)$ is the window function. An appropriate choice of window length is important to the calculation of short-term mean energy. If the length of window is

large, then E_n changes little over time, and the details will be neglected. On the contrary, if the length is short, then we could not get smooth energy features. Figure 4 shows short-term mean energy when the length of window is 200.

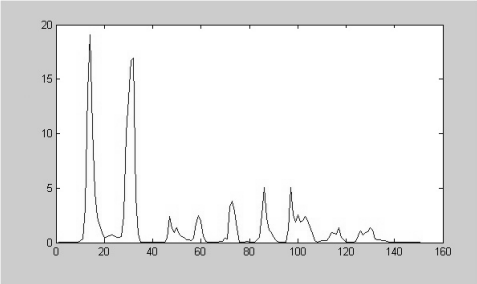

Figure 4 Short-term mean energy

Fundamental frequency of emotional speech reflects Quasi-periodicity of speech signals. The frequency of vocal cord vibration is called fundamental frequency. The reciprocal of fundamental frequency is fundamental cycle. We calculate fundamental frequency of each frame of speech signals. Fundamental frequency can be computed through time domain analysis, cepstrum domain analysis, or linear prediction analysis. We compute Fundamental frequency by time domain analysis. The procedures are as follows.

1. Conduct center clipping on speech signals. After pre-processing, the amplitudes at integer multiples of fundamental cycle are strengthened. Figure 5 shows the center clipping result of a period of speech signals.

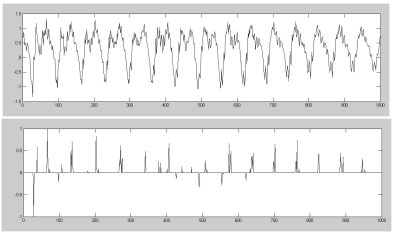

Figure 5 Center clipping

2. Compute short-term self-correlation function of each speech frame, which is shown in Figure 6. The formula of short-term self-correlation function is as follows.

$$r_n(k) = \sum_{m=n}^{n+(N-1)} [x(m)\omega(n-m)][x(m+k)\omega(n-(m+k))]. \quad (2)$$

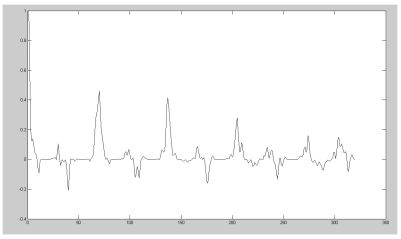

Figure 6 Short-term self-correlation function

3. Obtain fundamental cycle based on the first maximum of self-correlation function. The reciprocal of fundamental cycle is fundamental frequency.

4. Conduct smoothing on fundamental frequency curve to remove noise.

Formant frequencies can be acquired through Discrete Fourier Transform. The maxima of spectrum curve are thought as formant frequencies. We can also calculate formant frequencies by cepstrum domain analysis or linear prediction analysis.

3. Emotion recognition fusing facial expression and speech

We construct a multi-modal fusion emotion recognition framework based on Hidden Markov Models and Multilayer Perceptron.

3.1. Hidden Markov Models for facial expression and speech

Hidden Markov Models (HMMs) can be applied to solve time-related problems. If samples are feature vectors corresponding to continuous time, then they can be used to train a Hidden Markov Model. In this paper, feature vectors extracted from facial expression image sequences and emotional speech are vectors corresponding to continuous time. So we use them to train HMMs for all kinds of facial expressions and emotional speech.

The state set of an HMM is $S = s_i, i = 1, 2, ..., N$, where N is the number of states. The observation set is $V = v_i, i = 1, 2, ..., M$, where M is the number of observation vectors. Probability distribution at the initial time is $\pi = \pi_i$. State transition probability matrix is $A = a_{ij}$. a_{ij} is the state transition probability where state at this moment is i, and state at the next moment is j. State probability distribution matrix is $B = b_j(v_k)$. $b_j(v_k)$ is the probability where state is j and the observation vector is v_k .

In this paper, we apply left-to-right continuous HMMs. Observation vectors used to train HMMs are facial expressions and emotional speech features. The observation vec-

tors accord Gaussian mixture distribution. Figure 7 shows a four-state left-to-right HMM.

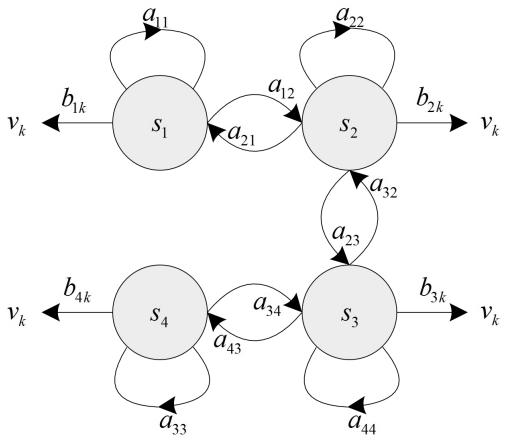

Figure 7 left-to-right Hidden Markov Model

We use Viterbi algorithm to train the parameters of HMMs. The iterative procedures of the training algorithm are as below.

1. Conduct uniform segmentation on state sequences corresponding to observation vectors of the training set.

2. Apply K-Means clustering algorithm to distribute Gaussian mixtures to observation vectors for all states.

3. Obtain probability density function of observation vectors for all states based on weights, mean vectors and covariance matrixes of Gaussian mixtures of the states.

4. Conduct Viterbi segmentation on state sequences corresponding to observation vectors of the training set.

5. Redistribute Gaussian mixtures to observation vectors for all states based on Euclidean Distance between observation vectors and mean vectors of Gaussian mixtures.

6. Repeat 3-5, until sum of maximums of $P(O, Q \mid \lambda)$ do not change any more.

For a feature vector sequence of video of audio, we calculate the conditional probabilities $P(O \mid \lambda)$ for all HMMs. HMM which has the biggest $P(O \mid \lambda)$ is the HMM corresponding to the affective state of feature vector. We maintain the conditional probabilities in order to use them in modal fusion. We apply Multi-layer Perceptron to fuse video and audio modalities on decision level, $P(O \mid \lambda)$ can be used as inputs of a Multi-layer Perceptron.

3.2. Multi-modal fusion based on Multi-layer Perceptron

Multi-layer Perceptron simulates neural organization. It reflects the structure and learning process of neural organization. The basic unit of Multi-layer Perceptron is neuron. Though structure and function of each neuron is simple, network made up of large numbers of neurons have complicated structure and strong function. Multi-layer Perceptron have learning ability and fault-tolerant performance. Like human braisn, Multi-layer Perceptron reflects the mutual impacts and restrictions among all sources, and it helps fuse information from all modalities. As a result, Multi-layer Perceptron is an appropriate method that can be used in multi-modal fusion issues.

The unit of Multi-layer Perceptron is neuron. Every neuron has several input connectors which turn outputs from the previous layer into inputs of the neuron. It also has several output connectors which deliver responses to the next layer. Figure 8 shows a neuron, where $y_j = f(\sum_{i=1}^{N} x_i \omega_{j,i} + bias_j)$.

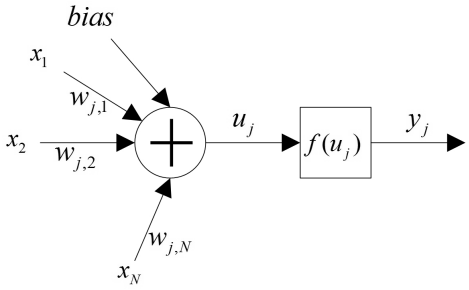

Figure 8 Structure of a neuron

Multi-layer Perceptron consists of an input layer, an output layer and several hidden layers. Each layer has several neurons which are connected with neurons of the previous layer and the next layer. Figure 9 shows the structure of a Multi-layer Perceptron. It has two neurons in the input layer, five neurons in the hidden layer, and two neurons in the output layer.

In order to train Multi-layer Perceptron which fuses facial expression and emotional speech, first of all, we extract emotional features from facial image sequences and speech signals. Features are used as observation vectors to train Hidden Markov Models of different emotions in different modalities. Calculate conditional probabilities $P(O \mid \lambda)$ for all HMMs. These conditional probabilities are used as inputs of Multi-layer Perceptron. Number of neurons in output layer is the number of affective states after modal fusion.

Inputs of Multi-layer Perceptron are delivered to the first hidden layer. Outputs of hidden layers are calculated based on weights and activation functions and conveyed to the next layers. So we need to train the weight matrix of Multi-layer Perceptron. Back-propagation learning algorithm is applied in the training process.

Test samples are facial expression image sequences and corresponding emotional speech signals. We extract ex-

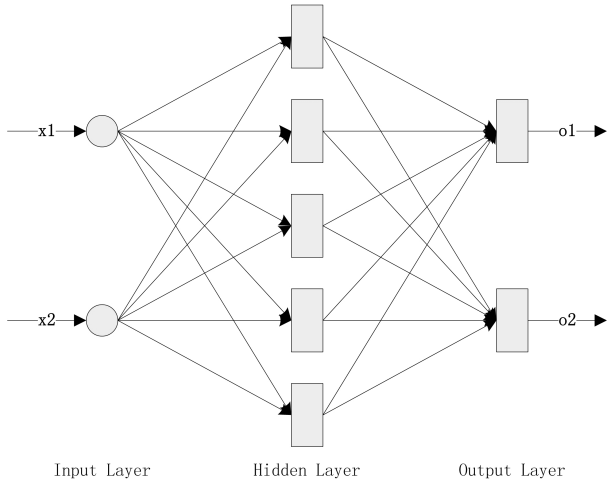

Figure 9 Structure of a neuron

pression features and speech features from test samples, and calculate conditional probabilities $P(O \mid \lambda)$ for all HMMs as inputs of Multi-layer Perceptron. Outputs of Multi-layer Perceptron are acquired based on weight matrix of the trained Multi-layer Perceptron. As each neuron in the output layer represent an affective state. The neuron which has the biggest output value corresponds to affective state of the test sample.

The number of hidden layers and the number of neurons in each layer have an effect on the training of Multi-layer Perceptron. Errors on the training set will decrease when the numbers increase. However, at the same time, network will learn the noise of training set. So pre-processing is always conducted on training data, such as Principal Component Analysis and so on. In the experiment, we apply Multi-layer Perceptron with two hidden layers. Number of neurons in each hidden layer is 20.

4. Experiments

In the experiments, we use Cohn-Kanade 2010 Facial Expression Database and Berlin Emotional Speech Database. Cohn-Kanade Database consists of facial expression image sequences of 123 volunteers. Each image is frontal face with constant light intensity and different facial expressions. Each sequence starts from neutral expression, and ends at the peak of expression. We use expression images in the database to train an Active Appearance Model. Facial feature points are detected and tracked on the model. Facial Animation Parameters are computed as observation vectors of HMMs based on motions of feature points.

Berlin Emotional Speech Database involves 433 pieces of speech. Each piece of speech lasts for 1-3 seconds. Contents of speech are ten German sentences which are read in six kinds of affective states by ten volunteers (five men and five women). We extract energy, fundamental frequency and formant frequencies to construct observation vectors of HMMs. Training samples of Multi-layer Perceptron are conditional probabilities $P(O \mid \lambda)$ for all HMMs. The trained Multi-layer Perceptron can be used to classify affective states of test samples. There are four affective states tested by the experiments, including happiness, sadness, anger and disgust. Table 2 is the number of samples of each affective state.

Table 2 Number of samples in each affective state

Happiness	Sadness	Anger	Disgust
62	28	45	36

After multi-modal fusion, the recognition rate of test samples reaches 91.2% through cross-validation. Mean recognition rate on uni-modality is only 69.3%. Table 3 shows the recognition results.

Table 3 Emotion recognition rates

	Happiness	Sadness	Anger	Disgust
Uni-modality	64.5%	89.3%	58.9%	75.0%
Multi-modality	95.2%	92.9%	93.3%	80.6%

5. Conclusions and future work

Multi-modal fusion emotion recognition is an important research field of pattern recognition. In this paper, we construct a multi-modal fusion emotion recognition framework based on Multi-layer Perceptron and Hidden Markov Models. Experiments show, the multi-modal approach has relatively high recognition rate. In the future, we will improve the extraction algorithms of emotional features. More features will be introduced into our algorithm for emotion classification.

Acknowledgement

This work is partially supported by the National 973 Foundation of China (No. 2013CB329304), the National 985 Foundation of China, National Science Foundation of China (No. 61222210), Major National Science and Technology Programs (No. 2009ZX09502) and the National 863 Program of China (No. 2013AA013204).

References

[1] P. Ekman, R. Davidson, The Nature of Emotion: Fundamental Questions. Oxford University Press, New York, (1994).

[2] Wang Xue-guang, Chen Shu-hong, An Improved Image Segmentation Algorithm Based on Two-Dimensional Otsu Method, Inf. Sci. Lett. 1 (2012) Vol. **1**, No. 2, 77-83.

[3] L. Franco, A. Treves, A neural network facial expression recognition system using unsupervised local processing. In Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. (2001), 628-632.

[4] Z. Ying, J. Tang, J. Li, et al., Support Vector Discriminant Analysis and Its Application to Facial Expression Recognition. Chinese Journal of Electronics, (2008), Vol. **36**, No.4, 725-730.

[5] J. Bassili, Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Journal of Personality and Social Psychology, (1979), Vol. **37**, No. 11, 2049-2058.

[6] N. Zhao, Y. Liu, Product Approximate Reasoning of Online Reviews Applying to Consumer Affective and Psychological Motives Research. Applied Mathematics & Information Sciences, (2011), Vol. 5, No. 2, 45-51.

[7] P. Ekman, W. Friesen, Facial Action Coding System. Consulting Psychologists Press, California, (1978).

[8] M. Yu, S. Li, Dynamic Facial Expression Recognition Based on Optical Flow. Microelectronics and computer, (2005), No. 7, 113-115.

[9] G. Yang, Z. Yu, Facial Expression Recognition Based on Improved Optical Flow Algorithm and HMM. Micro computer information, (2008), No. 1, 284-286.

[10] J. Yeh, M. Yen, Speech Recognition with Word Fragment Detection Using Prosody Features for Spontaneous Speech. Applied Mathematics & Information Sciences, (2012), Vol. **6**, No. 2, 669-675.

[11] L. Zhao, C. Jiang, C. Zou et al., A Study on Emotional Feature Analysis and Recognition in Speech. Chinese Journal of Electronics, (2004), Vol. **32**, No. 4, 606-609.

[12] Y. Lin, G. Wei, Speech emotion recognition based on HMM and SVM. In Proceedings of the Fourth International Conference on Machine Learning and Cybernetics. (2005), 4898-4901.

[13] J. Ye, M. Zhang, X. Gong, Speech emotion recognition based on MF-DFA. Computer engineering and applications, (2012), Vol. **48**, No. 18, 119-122.

[14] Z. Zeng, J. Tu, M. Brian, T.S. Huang, Audio-visual affective expression recognition through multistream fused HMM. IEEE Transactions on Multimedia, (2008), Vol. **10**, No. 4, 570-577.

[15] H. Go, K. Kwak, D. Lee et al., Emotion recognition from facial image and speech signal. In Proceedings of International Conference on Instrument and Control Engineers. (2003), 2890-2895.

[16] P. Lucey, J.F. Cohn, T. Kanade et al., The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of IEEE workshop on CVPR for Human Communicative Behavior Analysis. (2010), 94-101.

Chao Xu received his Ph.D at School of Computer Science and Technology, Tianjin University. He is currently a lecturer in Tianjin University. His research interests lie in Knowledge Management, Pattern Recognition, Security Software engineering, and Affective Computing.

Pufeng Du is currently a lecturer in Tianjin University. He received his Ph.D at Tsinghua University. His research interest includes machine learning theory, protein function prediction, RNA editing sites analysis and bioinformatics software development.

Zhiyong Feng received his Ph.D at Tianjin University. He is currently a professor in Tianjin University. His research interests lie in Knowledge Engineering, Services Computing, Security Software Engineering and Computer cognitive.

Zhaopeng Meng is currently a professor in Tianjin University. His research interests lie in CSCV-based Collaborative systems, Computer networks and applications, Computer distance education.

Tianyi Cao is a graduate student in School of Computer Science and Technology, Tianjin University. He received his Bachelor's degree from Hebei University of Technology in 2010. He has conducted research in pattern recognition, Security Software engineering, and Affective Computing.

Caichao Dong is a graduate student in School of Computer Science and Technology, Tianjin University. He received his Bachelor's degree from Yanshan University in 2010. His research field is image processing, pattern recognition and Affective Computing.