RESEARCH ARTICLE OPEN ACCESS

Research on Extracting Facial Image For Bimodal Emotion Recognition Based on Speech Signal and Facial Expression

Chuanghong Su*, Wei Zhang**,Tianhua Feng***

*(School of Information Science and Technology, Jinan University, Guangzhou, China)

--************************----------------------------------

Abstract:

On bimodal emotion recognition based on speech signal and facial expression, a method is proposed to extract the corresponding image farme combining speech signal frame on the timeline. The method overcomes the problem which the key frame of static facial expressional image can't be pinpointed as the recognized object on video. It is also use to combine nonlinear features of speech signal to conduct the study of emotional recognition. Firstly, linear and nonlinear features of the speech signal are extracted as the speech emotion feature. After pre-processing speech signal frames and adds window according to the overall frames of video, the mean of the absolute value of all framing signal sampling points' amplitude is computed, and ten images,which correspond to ten largest absolute values of speech signal, can be selected. And then, ten images' feature vectors are extracted and their mean is computed as last feature vector of facial expression of video. Finally, the support vector machine classifier is used to recognize emotion through feature-level fusion.Experiments show that the extracted facial features can better reflect the emotional state, and the recognition effect is achieved better though fusion of speech signal features.

Keywords — **Bimodal Emotion Recognition, Nonlinear Features of Speech Signal, Extraction of Facial Images of Video**

--************************----------------------------------

I. Summarize

In 1971, American psychologists, Ekman and Friesen, first proposed six basic emotions: anger, sadness, joy, fear, disgust and surprise[1].With the development of science and technology, the research about emotion recognition has become an important research in the field of machine learning. Emotion recognition research mainly makes use of the body's biological features and physiological signal feature, such as voice, facial expressions, blood pressure, heart signals, pulse and etc, to obtain for the experimental research of emotional state of experiment object. At present, emotional identification, based on emotional behavior at home and abroad, can be divided into single-modal emotional recognition and multi-modal emotional recognition. The single-modal emotional recognition is identified by single feature, and there are limitations in the recognition of the research process. By fusing multiple features, multi-modal emotional recognition increases the complexity of feature recognition and becomes more reliable and accurate.

Speech recognition and facial expression recognition are two important research directions of emotional recognition research, and bimodal emotional recognition which combines them also becomes one of hot spots of the experiment research. Cheng-Yao Chen and Yue-Kai Huang[2] demonstrated that the bimodal emotional features, which fuse speech and facial expression, are more effective than single-modal emotional features. Yao hui[3] used the nonlinear features of speech signals to identify four emotion speech, and points out the non-linear speech features have some effect on the speech emotional recognition. Huang et al.[4] carried out the bimodal emotional recognition by extracting the facial geometric features, phonological features and acoustic features, and also achieved good results. Although the bimodal emotional recognition of speech and facial expression has achieved good results, but the experiment process still has some problems, such as emotion databases which are used in research are very difficult to get unified use, and some databases which are known is mostly not made public. In addition, which the feature extraction of speech signals is still based on the extraction of linear features, nonlinear features of the speech signal are always ignored. To extract the audio and video facial expression image, some methods based on dynamic image information are used, such as optical flow method[5] and difference image method[6]. Those methods didn't combine voice to video image in extracting key frames of facial expressions.

In this paper, the emotional recognition study was carried out on eNTERFACE'05 emotion database[7]. The generation of speech is a complicated nonlinear

process[8], so nonlinear features of the speech signal are extracted as the test of the phonetic feature. On emotional recognition of facial expressions of video, the static image features of video are extracted with the

amplitude of speech signal and the corresponding average amount of image features is calculated as a new feature extraction. It avoids the question that the selection of key frame of single-modal video images is uncertain and subjective. And with the combination of voice signal amplitude changes,a new method is put forward to extract video static facial expression of image frames, which contributes to effective development in training and recognition of the late image features .

II. Nonlinear Features of Emotional Speech

For the nonlinear features of emotional speech, five kinds of features is selected to experiment, including hurst index, Lempel_Ziv complexity, lyapunov index, approximate entropy and box dimension. The detailed introduction is as follows:

A. Hurst Index

Hurst index (hereinafter referred to as "H") was designed by H. E. Hurst, a British hydrologist in the study of the relationship between water flow and the storage capacity in reservoir. The rescaled range analysis method(R/S)[9] is used to calculate the *H* of the emotional speech. In above method, discrete speech time series is extracted and divided into *N* adjacent subsequences *u* according to equal length *n*. Then the cumulative deviation z_u and standard deviation S_u of each sequence are calculated, and we also can calculate the range of z_u *as following.*

$$R_u = \max z_u - \min z_u \qquad (1)$$

And then the rescaled range R_u/S_u of each subsequence can be obtained, and finally we can get the calculation formula of the *H* as follows.

$$(R/S)_n = Kn^H \qquad (2)$$

Among them, (R/S)n is the mean of the values of all subsequences, K is constant, and H is the Hurst index. The features of the Hurst index can reflect the correlation difference of emotions around the time. That is to say, different emotional states can be distinguished by different Hurst exponents.

B. Lempel_Ziv Complexity

As a nonlinear analysis method, Lempel_Ziv complexity (hereinafter referred to as "LZC") is one of the important parameters of analyzing time series due to its small computation and easy implementation. LZC reflects that the time series emerge the rate of new model with the increase of time length[10] and shows the randomness of time series. According to LZC algorithm[10], assume the length of a time series is *n*, and get the number $c(n)$ of diverse substring in the reconstructed sequence $\{S(i)\}$. When *n* tend to infinity, $c(n)$will tend to $n/\log_L n$. *L* is expressed as segment number which is divided by the original sequence though a coarse graining preprocessing. Finally, the formula of the normalized LZC is as follows.

$$C = \frac{c(n)\log_L n}{n} \qquad (3)$$

Since LZC reflects the randomness of time series and different emotional speech signals is different, LZC can be used as the feature of emotional speech.

C. Largest Lyapunov Exponent

The lyapunov exponent is an important parameter to measure the divergence rate of average index in the neighboring trajectories of the phase space. The Largest Lyapunov Exponent (hereinafter referred to as "LLE"), reflects the convergence rate or the divergence rate of the orbit. And the value of LLE is indicated as when $\lambda_1 > 0$, the orbital divergence rate and chaos degree also increase as λ_1 increase. For the calculations of LLE, compared to other methods, the algorithm of small data quantity[11], put forward by Rosenstein et al., has good robustness for the selection of embedding dimension, time delay and average period. Therefore this article uses the algorithm of small data quantity to solve the problem. Its main calculation formula is as follows.

$$y(n) = \frac{1}{\Delta t}\langle \ln d_m(n)\rangle \qquad (4)$$

Among them, Δt represents the period of time series. $d_m(n)$ represents the distance of m'th similar points after *n* time step in the basic orbit.$\langle * \rangle$ denotes the mean value of all the values of *m* in the symbol.

D. Approximate Entropy

As a nonlinear dynamic parameter, approximate entropy (hereinafter referred to as "ApEn") doesn't require large amount of data and has strong anti-noise and anti-interference ability[12]. ApEn mainly focuses on the complexity of the process of time series from a statistical point of view, reflecting the self-similarity of time series in patterns. The algorithm of ApEn[12] assumes that the time series, whose length is *N*, is $\{u(i), i = 1,2,\Lambda\ , N\}$ $\{u(i), i = 1,2,\Lambda\ , N\}$. Then the new dimension vector is reconstructed, and its vector form is as follows:

$$X_i = \{u(i), u(i+1), \Lambda\ , u(i+m-1)\} \qquad (5)$$

And it also calculates the relative Euclidean distance d_{ij} between any vector X_i and the remaining vector X_j, and counts the number nd of $d_{ij} \le r \times SD$, in which *r* is given threshold, and *SD* is the standard value of the sequence. The calculation formula can be got as follow:

$$C_i^m(i) = nd/(N-m) \qquad (6)$$

Then, we can take the log of $C_i^m(i)$, and we take the mean of $C_i^m(i)$.

$$\phi^m(r) = \sum_{i=1}^{N-m+1} \ln C_i^m(r) \Big/ (N-m+1) \qquad (7)$$

Finally, the formula of ApEn can be got as follow.

$$ApEn = \sum_{N\to\infty} \left|\phi^m - \phi^{m+1}\right| \qquad (8)$$

ApEn is widely used in bioelectrical signal, mechanical fault signal and current signal. Therefore, it is extracted as the nonlinear feature of emotional speech.

E. Box-counting Dimension

Box-counting dimension is a method to describe fractal theory in chaotic signals. Compared with other fractal dimensions, it has many advantages such as simple calculation, easy method to implement and convenient application on experiment. To calculate this dimension for a fractal S, we can imagine this fractal lying on an evenly-spaced grid, and count how many boxes are required to cover the set. The box-counting dimension is calculated by seeing how this number changes as we make the grid finer by applying a box-counting algorithm. In order to show the box dimension D_b of the subsubset F in the N-dimensional Euclidean space, we can use the formula as follow[13].

$$D_b = \lim_{\sigma \to 0} \frac{\log N_\sigma(F)}{\log(1/\sigma)} \qquad (9)$$

In the formula, $N_\sigma(F)$ is the number of smallest cubes whose length of each side is σ and which are covered by area F. The auto-correlation degree of speech signal increase with the decrease of the box dimension. Therefore, as one of nonlinear features of speech signal, box-counting dimension helps us to study the emotional speech recognition better.

III. the Extraction and Feature Processing of Facial Expression Image in Video

In video, the facial expression recognition mainly includes the movement feature recognition and static facial image recognition. The former focuses on the change of the movement information of facial expression, while the latter carry out the emotional recognition by selecting the key frame of the effective facial expression image in video.This paper mainly carries on the experiment through the latter.

A. the Extraction of Facial Expression Feature

For the extraction of static facial image feature, the method of improved local binary model and Gabor wavelet transform are used to extract features.

1. Local Binary Pattern

Local Binary Pattern(hereinafter referred to as "LBP"), put forward by t. Ojala et al. based on the operator of texture feature about image gray value, reflects the relationship between the image pixel of a position and the pixel in the domain. This paper uses the modified LBP operator of Uniform model[14] to extract the image feature. The operator, which can be expressed as $LBP_{P,R}^{u2}$, is represented by the following formula.

$$LBP_{P,R}^{u2} = \begin{cases} \sum_{i=0}^{P-1} s(g_i - g_c), & if \quad U(LBP_{P,R}) \le 2 \\ P+1, & otherwise \end{cases} \qquad (10)$$

Among above formula,

$$U(LBP_{P,R}) = |s(g_{p-1} - g_c) - s(g_i - g_c)| + \sum_{i=1}^{p-1} |s(g_i - g_c) - s(g_{i-1} - g_c)| \qquad 11)$$

$$s(x) = \begin{cases} 1, x \ge 0 \\ 0, x < 0 \end{cases} \qquad (12)$$

g_c represents the gray value of the center pixel, and $g_p(p = 0,1,\mathrm{K}\ , P-1)$ is the gray value of the pixel of the sample point in the circle. When using LBP operator to extract the facial expression feature, firstly we use MATLAB toolbox of Face Part Detection to extract the facial expressions of two eyes and mouth which can perform the emotion well. According to the method of 59-bin LBP operator[15], after extracting LBP histogram of three areas and connecting them in order, we can get LBP feature vector of the facial expression image, whose dimension is $177(59 \times 3)$.

2. Gabor Wavelet Transform

Gabor wavelet transform can effectively extract local structure information in both spatial domain and frequency domain. And the two-dimensional Gabor wavelet can show the texture feature of the image relatively clearly, so it is widely used in facial expression recognition. For a grayscale image $I(x, y)$, its Gabor wavelet transform formula is as follow.

$$O_{\mu,\nu}(x, y) = \left\| I(x, y) * \varphi_{\mu,\nu} \right\| \qquad (13)$$

Among above formula, $\varphi_{\mu,\nu}$ represents the kernel function of the two-dimensional Gabor wavelet, and μ and ν respectively represent the direction and scale of the Gabor filter. In order to reflect the texture feature of image better, we select 40-filter combination with 8 scale and 4 direction. Because the extracted feature dimension, based on Gabor wavelet transform, is big, we should ensure the treatment efficiency of subsequent recognition. Therefore, we need to take an average of Gabor feature vector and treat them with feature dimension reduction. this article[16] uses the method of principal component analysis (hereinafter referred to as "PCA") to reduce dimension and retain 90% of the original information.

B. the Extraction of the Image Frame of Static Facial Expression

Generally speaking, when people express emotion, the amplitude of speech signal is also enhanced, and there are better emotional expression on their facial expressions. Based on the relationship between emotional voice and image frames, a new idea of the feature processing of the static image frame of facial expression is presented as follows:

1) In the experiment, we extract and preprocess the speech signal. And the preprocessing process mainly includes the endpoint detection, the preaccentuation of signal filtering and the amplitude normalization. In this paper, the endpoint detection is carried out mainly through the method of energy-entropy ratio, and the preemphasis uses the high-pass filter. Then in the emotion database of eNTERFACE '05, the speech signal of angry emotion in the video s1_an_1 is extracted and pretreated. Figure 1 shows the contrast between the original speech signal and the time domain waveform of the preprocessing speech signal.

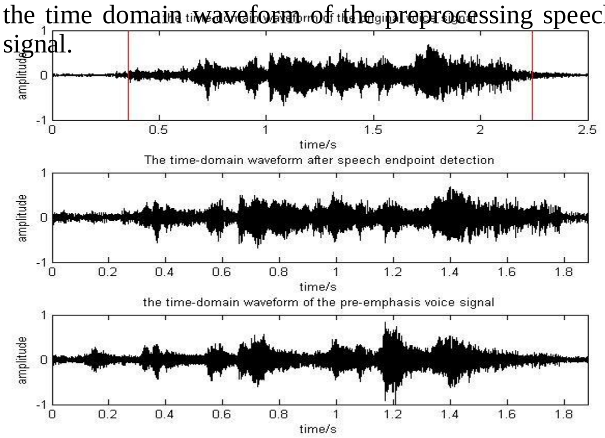

Fig. 1 Pre-processing of Speech Signal

2) After receiving the pre-processed emotional speech, we can frames and adds window to it. In order to contract

speech signal frame and image frame, assume that the number of inage frames per second is n, and the speech signal is framed and windowed in accordance with the n frames per second. At the same time, the hanning window function is added to the signal and the sampling frequency of speech signal is set at 16000Hz. Assume that the frame shift inc of the voice signal is equal to one third of frame length $wlen$,that is to say, $inc = wlen/3$. And then the calculation formula of frame length is shown as follows.

$$wlen \times n - (n-1) \times inc = 16000 \qquad (14)$$

After transforming the formula, we can get:

$$wlen = 48000/(2n+1) \qquad (15)$$

In this paper, the video image frame of the eNTERFACE '05 database is set at 25 frames per second, and the preprocessed speech signal frames and adds window at 25 frames per second.

3) After framing and adding window, assume the framing speech array is represented as $\{f(i)\}$, in which N is denoted as the total frame number, and $f(i)$ is time sequencee of speech signal of the i'th frame, which is expressed as follow.

$$f(i) = \{u_i(j), j = 1,2,\Lambda\ , wlen\} \qquad (16)$$

We take the mean of the absolute value of all the amplitudes $u_i(j)$ in each of the frame signals $f(i)$, and we use $s(i)$ to represent the mean. The formula is following.

$$s(i) = \sum_{j=1}^{wlen} |u_i(j)| \qquad (17)$$

After calculating the mean value of the absolute value of the amplitude, the formula is obtained as follow.

$$x = \{s(i), i = 1,2,..., N\} \qquad (18)$$

On the timeline, there is a one-to-one correspondence between each element of the sequence and the image frames of video. The sequence x is rearranged in the order from large to small. At the same time, rearranging the sequence of corresponding image frames, we can get the new image frame sequence as follow.

$$y = \{I(i), i = 1,2,..., N\} \qquad (19)$$

$I(i)$ represents the image frame. Figure 2 is the first five facial expression grayscale images of the sequence y (from top to bottom: surprise, anger, happiness, sadness, disgust and fear).

IV. Process and Analysis of the Experiment

A. Emotional Database of Video

We select the eNTERFACE'05 multi-modal emotional database, in which the experimental objects include 43 English-speaking people from different countries. The database collects emotional data by inducing emotion and contains six basic emotions, such as surprise, anger, happiness, sadness, disgust and fear. Each video contains five different sentences.The number of total experimental data is 1290.

B. Feature Vector of Emotional Speech

The feature extraction of emotional speech includes linear feature extraction and non-linear feature extraction. In this paper, four common linear features including fundamental frequency, resonance peak, short-time energy and Mel-frequency cepstral coefficient, are selected in the experiment. The total dimension of feature vector is 203. Before extracting the nonlinear features of speech, the speech signal should be preprocessed. The pretreatment mainly includes endpoint detection, pre-accentuation, amplitude normalization, framing and adding window. Among them, the endpoint detection adopts the energy-entropy-ratio method, the pre-accentuation uses the high-pass filter, whose coefficient is set to 0.99. The hanning window function are adopted to add window and the frame length is set at 512, and the frame shift is set at 256. After preprocessing the speech signal and extracting five nonlinear features including the hurst index, Lempel_Ziv complexity, Lyapunov index, box dimension and approximate entropy, and their first order difference, the maximum, minimum, mean, median, variance, skewness, and kurtosis of those features can be calculated respectively. The total feature dimension of the emotional speech is 273.

C. Construction of Feature Vectors of Facial Expression

Before extracting the feature of expression image, we should proceed face detection and extraction of the original image. We uses the Face Part Detection toolbox in MATLAB to detect and locate the static image and cut the image to 160×200 pixel. Figure 2 shows the gray images of six emotion in image sequence y(from top to bottom: surprise, anger, happiness, sadness, disgust, and fear). It can be seen that the first 5 images of the image sequence y, which is extracted by the method of 3.2, can represent their emotional state very well. Figure 3 shows the comparison between the images of the first 10 images and the last 10 images in the video s1_ha_1 which represent happy emotion. The stronger the amplitude of speech signal is, the better the corresponding facial expression image reflects the character's emotional state. on the contrary, the weaker the amplitude of speech signal is, the less obvious the corresponding facial expression image reflects the character's emotional state.

Therefore, in order to prevent single facial expression image can't completely represent character's emotional state in the video, we select the first 10 images of the image sequence y, which can represent emotional state, and respectively extract LBP feature vector and Gabor feature vector. After calculating the average of those 10 images as the average feature vector of the facial expression in the video, we can use PCA method to reduce dimension. Finally, the vector dimension of image is 257.

Fig. 2 Facial Expression Images of Six Emotions

Fig. 3 Facial Expression Images In the Video of s1_ha_1

D. Bimodal feature fusion of speech and facial expression

After extracting feature vector of speech emotion signal and facial expression, we need to fuse different modal feature vector. The fusion methods mainly include feature-layer fusion and decision-level fusion[17]. We select the method of feature-layer fusion to carry out the experiment. Feature-layer fusion shows that the feature vector of speech and facial expression are extracted and connected to form a new feature vector. Then by support SVM classifier,we can use the 10-fold cross-validation way to recognize emotion. The fusion method is easy to implement and also compresses objective information. Figure 4 shows the bimodal emotion recognition process based on feature-layer fusion.

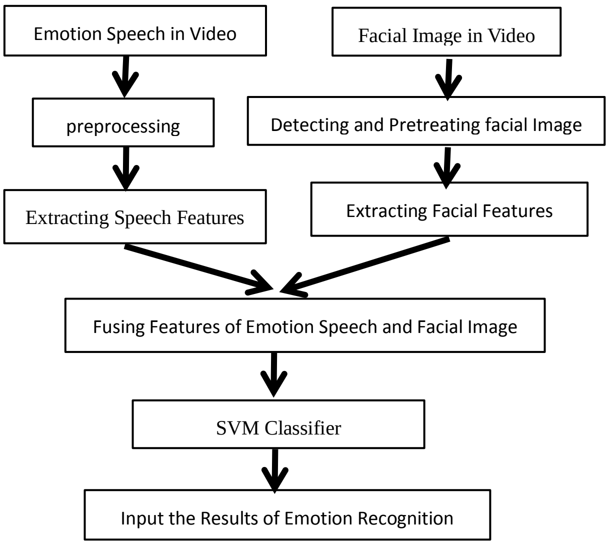

Fig. 4 Bimodal Emotional Recognition Based on Feature-level Fusion

E. the Experiment Process and Analysis

By using the SVM classifier to identify the linear and non-linear features of speech signals in the eNTERFACE '05 database,we can get the recognition sutuation of the six kinds of emotions is shown in table 1.

As can be seen from table 1, the nonlinear feature recognition rate of the three kinds of speech emotion of happiness, sadness and fear, is higher than the linear feature-recognition rate. They are increased by 2.79%, 4.74% and 7.24% respectively. By the combination with linear and nonlinear feature, the speech recognition rate has a certain improvemen. And compared with the linear feature recognition rate and the non-linear feature recognition rate, the average recognition rate increases by 4.92% and 7.08% respectively. It can be seen that the introduction of nonlinear features can effectively distinguish the speech emotion. Although in general the average recognition rate of nonlinear feature is less than the average recognition rate of linear feature, for some emotion state, such as happiness, sadness and fear, the recognition effect of nonlinear features better than that of linear feature. At the same time, speech emotion recognition which fuses linear feature and nonlinear feature can represent better recognition effect.

Table 1
Comparison of Recognition Rates of Different Types of Features of Emotional Speech(%)

	Linear Feature	Nonlinear Feature	Linear and Nonlinear Feature
Surprise	68.63	53.46	72.63
Anger	78.04	73.87	83.56
Happiness	68.74	71.53	73.71
Sadness	75.01	79.75	81.61
Disgust	63.85	55.45	64.89
Fear	58.39	65.63	65.78
Average	68.78	66.62	73.70

Table 2 shows the recognition rate of the static face expression in video database by using the SVM classifier. It can be seen from table 2 that by combining the Gabor feature and LBP feature of facial expression, the average recognition rate is generally improved and reaches 63.92%. But relative to the emotion speech recognition, the recognition effect of facial expression is relatively poor, so using single features of facial expression recognition can't achieve the ideal effect.

Table 2
Comparison of Recognition Rates of Different Types of Features of Facial Expression (%)

	Gabor Features	LBP Features	Gabor and LBP Features
Surprise	64.34	42.52	65.01
Anger	56.92	32.46	64.75
Happiness	73.80	59.74	76.46
Sadness	67.64	51.81	64.99
Disgust	64.46	48.02	62.07
Fear	43.36	32.99	50.22
Average	61.75	44.59	63.92

Table 3 is an emotional comparison between the two modal emotion recognition and the single-mode emotion recognition with the fusion of speech emotion and facial expression features.From table 3, it can be seen that the recognition rate of dual modal emotion recognition, which combines emotional speech features and facial expression features, has better recognition effect than emotion recognition of single mode.

Table 3
Comparison of Recognition Rates of Single Emotional Feature and Fusion Feature(%)

	Speech Features	Facial Features	Fusion Features
Surprise	72.63	65.01	78.56
Anger	83.56	64.75	87.31
Happiness	75.71	72.46	80.34
Sadness	81.61	64.99	85.75
Disgust	64.89	62.07	69.75
Fear	65.78	50.22	72.56
Average	74.03	63.25	79.05

REFERENCES:

[1] Ekman, P., Friesen, W.V.: Constants Across Cultures in the Face and Emotion[J]. Journal of Personality and Social Psychology 17(2), 124–129 (1971).

[2] Chen C Y, Huang Y K, Cook P. Visual/Acoustic emotion recognition[C]//Multimedia and Expo,2005. ICME 2005. IEEE International Conference on. IEEE. 2005: 1468-1471.

[3] Yao H, Sun Y, Zhang X Y. Research on nonlinear dynamics features of emotional speech[J]. Journal of Xidian University. 2016, 43(05): 167-172.

[4] Huang Y M, Zhang G B, Liu H B, Da F P. Emotion System on New Double-Mode Fusion Algorithm[J]. Journal of Tianjin University, 2010, 43(12): 1067-1072.

[5] Wang Bin, Weng Zhengkui, Wang Kun, etal. Real-time Tracking Method for Eye Feature Points Based on Lucas-Kanada Optical Flow Algorithm[J]. Computer Engineering, 2015, 41(7): 244-249.

[6] Wang Chuang. On the Facial Expression Recognition System Based on Dynamic Sequence Feature[D]. Jilin University, 2013.

[7] Eyben F, Wollmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor[C]// Proceedings of the international conference on Multimedia. ACM, 2010: 1459-1462.

[8] BANBROOK M, MCLAUGHLIN S, MANN I. Speech Characterization and Synthesis by Nonlinear Methods[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(1): 1-17.

[9] Craciun D, Isvoran A, Avram N M. Long range correlation of hydrophilicity and flexibility along the calcium binding protein chains[J]. Physica A: Statistical Mechanics and its Application, 2009, 388(21): 4609-4618.

[10] Lempel A, Ziv J. On the complexity of Finite sequence[J]. IEEE Trans Inf Theory, 1976, IT22: 75-81.

[11] Zhang X D, Liu X, Zhao P D. Methods for Calculating the Main-Axis Lyapunov Exponents of A Type of Chaotic Systems with Delay[J]. Acta Physica Sinica, 2009, 58(7): 4415-4420.

[12] Zhang Yan-yan. The Theory of Approximate Entropy and its Application[J]. Chinese Journal of Medical Physics, 2009, 26(06): 1543-1546.

[13] Wei G, Lu Y Q, Ouyang J Z, The Theory of Chaos and Fractal and Speech Signal Processing[J]. Acta Electronica Sinica, 1996, 24(1): 34-38.

[14] T. Ojala, M. Pietikinen, T. M enp. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE, Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[15] C. Shan, S. Gong, P. McOwan. Robust facial expression recognition using local binary Pattern[C]. IEEE International Conference on Image Processing(ICIP), Genoa, 2005, 370-373.

[16] G. Donato, M. Bartlett, J. Hager, etal Classifying facial actions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(10): 974-989.

[17] A. Jaimes, M. Hatef, R. P. Duin, etal. On combining classifiers[J]. Computer and Image Understanding , 2007, 108(1-2): 116-134.