Bimodal Emotion Recognition Based on Speech Signals and Facial Expression

Binbin Tu and Fengqin Yu

School of Internet of Things Engineering, Jiangnan University,
Wuxi 214122 China
tbbice@126.com

Abstract. Voice signals and facial expression changes are synchronized under the different emotions, the recognition algorithm based audio-visual feature fusion is proposed to identify emotional states more accurately. Prosodic features were extracted for speech emotional features, and local Gabor binary patterns were adopted for facial expression features. Two types of features were modeled with SVM respectively to obtain the probabilities of anger, disgust fear, happiness, sadness and surprise, and then fused the probabilities to gain the final decision. Simulation results demonstrate that the average recognition rates of the single modal classifier based on speech signals and based on facial expression reach 60% and 57% respectively, while the multimodal classifier with the feature fusion of speech signals and facial expression achieves 72%.

Keywords: speech emotion recognition, facial expression, local Gabor binary patterns, support vector machine, fusion.

1 Introduction

The voice contains the rich information of emotion, while the face as the most important external feature can transmit much non-verbal information to enhance, understand or express emotion [1]. And psychologist Mehrabian gave a emotional formula as feelings show = 7%utterance+ 38%voice+ 55%countenance. In recent years, the emotion recognition has developed towards multi-modal. Professor Picard of MIT Media Lab extracted the multi-modal features based on facial expressions, head gestures and so on to monitor real-time emotional states of the students in the learning [2].

The emotional information of voice is contained in the changes of the acoustic parameters including prosodic parameters and spectral parameters [3]. LPCC reflects the characteristics of channel physiological structure, while MFCC reflects the nonlinearity of people's auditory frequency [4]. Local Binary Pattern (LBP) proposed by T. Ojala is an effective texture description operator and can measure and extract the texture information in the local neighborhood of gray-scale image [5]. T. Ahone firstly introduced the theory of LBP to describe the face image areas face, divided face images into several regions, extracted the LBP texture features from them and achieved good results [6]. And Gabor wavelet transform can describe the multi-scale, multi-direction local features of the gray-scale image, imitate the contours of single-cell receptive in the cerebral cortex and capture the prominent visual properties [7].

Y. Wang and T. Li (Eds.): Foundations of Intelligent Systems, AISC 122, pp. 691–696.
springerlink.com

Therefore, Local Gabor Binary Patterns (LGBP) can extract multi-directional, multi-scale local image features using Gabor transform, and then encode these features with LBP operator to effectively distinguish the different facial expression images.

The remainder of this paper is organized as follows. Section 2 describes the algorithm theory of LGBP and the fusion rules of classifiers. Section 3 presents simulation steps of extracting features from bimodal emotion signals and recognizing. Section 4 obtains the results based on speech signals and facial expression. Conclusions and discussions are given in section 5.

2 Algorithm Theory

2.1 LBP

LBP is originally used for texture analysis problems, which is successfully applied to the face recognition area in recent years. The principle of LBP is to calculate the binary sequence in the gray varying information of each pixel in the image, then encode the relation of the binary sequence to form LBP. So LBP features have much gray variation information. A pixel in an image denoted as $g(x_c, y_c)$,the original LBP operator labels the pixels of an image by threshold the 3×3 neighborhood of each pixel $g_0, g_1, \cdots, g_7$ with the center value g_c ,defining the texture of the local area as $T = t(g_c, g_0, g_1, \cdots, g_7)$,then binary processing as follows:

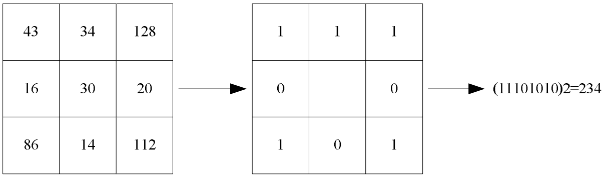
$$T \approx t\left(s\left(g_0 - g_c\right), s\left(g_1 - g_c\right), \cdots, g\left(g_7 - g_c\right)\right) \tag{1}$$

Where $s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$ .Describing the spatial structure of local texture characteristics of LBP expressed as

$$LBP_{8,1}\left(x_c, y_c\right) = \sum_{i=0}^{7} s\left(g_i - g_c\right)2^i \tag{2}$$

Where $LBP_{8,1}$ is LBP operator based on a circularly symmetric neighbor set of 8 members on a circle of radius 1, the LBP operator in Figure 1, a happy face image corresponding to the coded image shown in Figure 2.

With the increasing of the sampling points, the type of binary mode to increase dramatically, while the number of texture features dimension is larger, it is not conducive to classify. So T.Ojala proposed "uniform" patterns to improve the original LBP model, when the LBP corresponding to the binary number which changes from 0 to 1 or from 1 to 0 ,and its changes are no more than two times, U value is at most 2,



43	34	128
16	30	20
86	14	112

1	1	1
0		0
1	0	1

Fig. 1. LBP operator

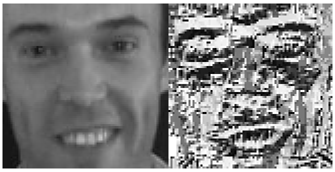


Fig. 2. A happy face image corresponding to the LBP image

denoted as $P_{8,1}^{u2}$,for example, patterns 00000000, 11111111, 11000011 are "uniform" patterns, defined as:

$$P_{8,1}^{u2} = \begin{cases} \sum_{i=0}^{7} s(g_i - g_c)2^i, U \le 2 \\ 8+1, \qquad otherwise \end{cases} \tag{3}$$

When using the (8, 1) field, the "uniform" patterns account for 90% of all patterns, which can effectively describe most of the image texture features.and reduce the number of features greatly. Therefore, this paper takes (8, 1) field and "uniform" patterns of facial expression features.

2.2 LGBP Operator

LGBP operator is to combine the amplitude of the Gabor wavelet features and LBP encoding. Gabor features are widely used in visual understanding information, the transform coefficients have well visual characteristics and are sensitive to the edge of the images. Gabor filters facilitate to adjust the direction, the base frequency bandwidth, and have a higher resolving power in time and frequency domain, two-dimensional Gabor wavelet kernel function is given:

$$\psi_{\mu,v}(z) = \frac{\left\|k_{\mu,v}\right\|}{\sigma^2} e^{\left(-\left\|k_{\mu,v}\right\|^2 \left\|z\right\|^2 / 2\sigma^2\right)} \left[e^{ik_{\mu,v}z} - e^{-\sigma^2/2}\right] \tag{4}$$

Where $z = (x, y)$ is image pixel, norm operation denotes $\|\bullet\|$, μ and v represent the direction and scale of Gabor filters respectively. $k_{\mu,v} = k_v e^{i\varphi_\mu}$, $k_v = 2^{-(v+2)/2}\pi$, $\varphi_\mu = \pi\mu / k$, k represent total number of wavelets directions, k_v is kernel frequency, and σ is wavelet filter bandwidth.

Gabor features of face images are gained by the convolution of facial images and Gabor filters, $f(x, y)$ is the gray distribution of facial images, and Gabor features as:

$$G(x, y, \mu, v) = f(x, y) * \psi_{\mu,v}(z) \tag{5}$$

Eight-direction and five-scale Gabor filters are adopted in this paper, $\mu = 0,1,\cdots, 7$, $v = 0,1,2,3,4$. LGBP is gained by the formula (2) and (5):

$$LGBP = \sum_{i=0}^{7} s\left(G_i(x, y, \mu, v) - G_c(x, y, \mu, v)\right)2^i \tag{6}$$

2.3 Fusion Rule

Kittler proposed the theoretical framework based on minimum error rate Bayesian classifier[8].Considering a pattern recognition problem where pattern Z is to be assigned to one of the C possible emotion classes $\Omega = \{\omega_1, \omega_2 \cdots \omega_c\}$, extracted R group features $f_1, f_2 \cdots f_R$ from a sample, we have R classifiers, and each classifier corresponding to the feature f_R , R group features are modeled by the SVM, then gained posterior probabilities $P(\omega_i \mid f_1) \sim P(\omega_i \mid f_R)$ of the i-th emotion. Take the maximum of posterior probabilities of the corresponding categories as the fusion result:

$$r(Z) = \arg\max F\left[P(\omega_i \mid f_1), \cdots, P(\omega_i \mid f_R)\right] \tag{7}$$

Function F represent fusion rule, including sum rule (8) and product rule (9).

$$F\left[P(\omega_i \mid f_1), \cdots, P(\omega_i \mid f_R)\right] = \sum_{j=1}^{R} P(\omega_i \mid f_j) \tag{8}$$

$$F\left[P(\omega_i \mid f_1), \cdots, P(\omega_i \mid f_R)\right] = \prod_{j=1}^{R} P(\omega_i \mid f_j) \tag{9}$$

3 Simulation Steps

Emotion recognition based on speech and facial expression was carried out as follows:

(1) Pre-process the emotional speech samples, including pre-emphasis and window them into frames. Frame size was set to 256 points, and frame shift 128, while the hamming win was chosen;

(2) For each frame signal, compute its FFT to gain the energy spectrum which will then input into Mel filter group to extract 12-dimensional MFCC;

(3) Compute the average fundamental frequency, short-time average energy, and 12-dimensional LPCC of each frame signal;

(4) Read the facial expression images, and pre-process them with geometry normalize and grayscale normalize, each image was normalized into 123×123,and maintain the eyes, eyebrows, mouth of the face image in the same location;

(5) Make eight-direction, five-scale Gabor wavelet decomposition for the images above, and take $\mu = 0,1,\cdots$, 7 , $v = 0,1,2,3,4$, $\varphi_\mu = 0, \frac{1}{8}\pi ;\cdots \frac{7}{8}\pi$, $\sigma = 2\pi$. Then achieve LBP code to calculate the LGBP features of each sub-images;

(6) Input speech features which include fundamental frequency, energy, LPCC,MFCC and image features which contain the LGBP into SVM respectively, finally the recognition rates of six emotions were obtained;

(7) Get all emotions' posterior probability based on speech features and image features respectively;

(8) Fuse the posterior probability of two features according to sum and product rules, calculate the recognition rates corresponding to the six emotions again.

4 Simulation and Results

The eNTERFACE’05 audio-visual emotion database was adopted in this paper, including six emotions of happiness, sadness, surprise, anger, disgust and fear, containing 42 subjects, recorded by the people of 14 different nationalities. The speech and facial samples were cut from the video files by 16 kHz sampling rate, 16 bit sampling precision and 25 frames/sec image sampling rate, we selected 900 samples of voice and facial samples of five persons as the experimental test sample.

Extracted pitch frequency, short-time average energy, 12-dimensional LPCC and MFCC from speech signals, then using SVM to identify six emotions, the results is shown in Table 1,average recognition rate is 60% based on speech signals, disgust and sadness are the most confusing emotion, surprise recognition rate is highest.

LGBP features are extracted from facial expression images, six kinds of emotional recognition results are shown in Table 2, the average recognition rate is 57%, when only using facial expression to identify emotions, fear and surprise are the most confusing emotions, happiness achieve the highest recognition rate.

Table 1. The recognition rate of the six emotions based on speech signal (%)

emotion	anger	disgust	fear	happiness	sadness	surprise
anger	60	17	0	0	23	0
disgust	0	53	7	0	33	7
fear	27	0	40	13	0	20
happiness	20	0	0	60	0	20
sadness	0	13	20	0	67	0
surprise	0	0	13	7	0	80

Table 2. The recognition rate of the six emotions based on facial images (%)

emotion	anger	disgust	fear	happiness	sadness	surprise
anger	60	13	0	27	0	0
disgust	0	40	6	0	27	27
fear	0	0	53	0	13	33
happiness	20	0	0	80	0	0
sadness	7	13	0	13	67	0
surprise	27	0	20	0	13	40

Using sum and product rules to fuse the features of speech signal and facial expression, the results of sum rule are better in table 3, the average recognition rate reach 72%, which is improved significantly. Anger and sadness recognition rate increased to 80%, fear and surprise confusing error rate have been reduced.

Table 3. The recognition rate based on fusion of speech signal and facial images (%)

emotion	anger	disgust	fear	happiness	sadness	surprise
anger	80	0	0	13	7	0
disgust	0	53	20	0	27	0
fear	0	0	67	0	20	13
happiness	13	0	0	87	0	0
sadness	0	13	0	7	80	0
surprise	7	0	13	0	13	67

5 Conclusion

While most emotion recognition researches rely on single-mode data, speech and image features were both extracted for bimodal emotion recognition in this paper. With fusion of speech and facial image features, an effective emotion recognition approach based on two kind biological signals was implemented. As is shown in the experiments, recognition rate after feature fusing was improved significantly.

Acknowledgments. This paper was supported by the National Natural Science Foundation of China No.61075008. The authors would like to thank Martin who provided eNTERFACE'05 audio-visual emotion database.

References

1. Jinjing, X., Yiqiang, C., Junfa, L.: Multi-expression Facial Animation based on Speech Emotion Recognition. Journal of Computer-aided Design & Computer Graphics 20(4), 520–525 (2008)
2. Kapoor, A., Picard, R.W.: Multimodal Affect Recognition in Learning Environments. In: Proc. of the 13th Annual International Conference on Multimedia, Singapore, pp. 677–682 (2005)
3. Danning, J., Lianhong, C.: Speech Emotion Recognition using Acoustic Features. J. Tsinghua Univ (Sci. & Tech.) 46(1), 86–89 (2006)
4. Koolagudi, S.G., Nandy, S., Rao, K.S.: Spectral Features for Emotion Classification. In: 2009 IEEE International Advance Computing Conference, Patiala, pp. 1292–1296 (2009)
5. Ojala, T., Pietikainen, M., Maenpaa, T.: Multi-resolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
6. Ahonen, T., Hadid, A., Pietikainen, M.: Face Description with Local Binary Patterns: Application to Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12), 2037–2041 (2006)
7. Wenchao, Z., Shiguang, S., Hongming, Z.: Histogram Sequence of Local Gabor Binary Pattern for Face Description and Identification. Journal of Software 17(12), 2508–2517 (2006)
8. Kittler, J., Hatef, M., Duin, R.P.: On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)