Multimodal Emotion Recognition Integrating Affective Speech with Facial Expression

SHIQING ZHANG [1], XIAOHU WANG [2], GANG ZHANG [3], XIAOMING ZHAO [1*]
[1] Institute of Image Processing and Pattern Recognition
Taizhou University
Taizhou 318000
CHINA
tzczsq@163.com, tzxyzxm@163.com (Corresponding author*, X.M ZHAO)
[2] School of Electronic and Information Engineering
Hunan Institute of Technology
Hengyang 421002
CHINA
xhwang_china2008@163.com
[3]Bay Area Compliance Labs. Corp.
Shenzhen 518000
CHINA
tzxyzz@126.com

Abstract: - In recent years, emotion recognition has attracted extensive interest in signal processing, artificial intelligence and pattern recognition due to its potential applications to human-computer-interaction (HCI). Most previously published works in the field of emotion recognition devote to performing emotion recognition by using either affective speech or facial expression. However, Affective speech and facial expression are mainly two important ways of human emotion expression, as they are the most natural and efficient manners for human beings to communicate their emotions and intentions. In this paper, we aim to develop a multimodal emotion recognition system integrating affective speech with facial expression and investigate the performance of multimodal emotion recognition at the feature-level and at the decision-level. After extracting acoustic features and facial features related to human emotion expression, the popular support vector machines (SVM) classifier is employed to perform emotion classification. Experimental results on the benchmarking eNTERFACE’05 emotional database indicate that the given approach of multimodal emotion recognition integrating affective speech with facial expression obtains obviously superior performance to the single emotion recognition approach, i.e., speech emotion recognition or facial expression recognition. The best performance obtained by using the product rule at the decision-level fusion is up to 67.44%.
.

Key-Words: - Mutimodal emotion recognition, affective speech, facial expression, support vector machines, speech emotion recognition, facial expression recognition

1 Introduction

The traditional human-computer-interaction (HCI) system, where an user facing a computer interacts with it by using a mouse or a keyboard, are employed to stress the transmission of explicit information while neglecting implicit information related to the user’s changes in the affective states. Such interactions are hence usually regarded as incompetent, cold, and socially inept. To address this problem, a recently-emerged research field of “affective computing” [1], which aims to make computers recognize, express, model, communicate and respond to a user’s affective information, has attracted extensive attentions in signal processing, computer vision, pattern recognition, One of the most important motivations is that affective computing aims to enable HCI to be more human-like, more effective, and more efficient [2-6]. Especially, such computers endowed with a capability of affective computing is able to communicate with human users based on the detected affective information, rather than give a simple response to a user's commands.

Emotion transfers the psychological information of a person, since emotion is conveyed by various physiological changes, such as changes in heart-beating rate, sweating degree, blood pressure, etc. Emotion is expressed by affective speech, facial expression, body gesture and so on. Among them, affective speech and facial expression are mainly two important ways of human emotion expression, as they are the most natural and efficient manners for human beings to communicate their emotions and intentions.

During the past two decades, enormous efforts have been devoted to developing automatic emotion recognition systems. Most of previously published work on emotion recognition focused on recognizing human emotions by using a mono-modality, such as the single facial expression or the single affective speech. The former is called facial expression recognition, and the latter is called speech emotion recognition. Motivated by very little work done on multimodal emotion recognition, in this paper we aim to investigate the performance of multimodal emotion recognition integrating affective speech with facial expression both at the feature-level and at the decision-level. To verify the performance of multimodal emotion recognition, the popular eNTERFACE'05 multimodal emotional database [7] is employed to perform multimodal emotion recognition experiments.

The remainder of this paper is structured as follows. A flowchart of a multimodal emotion recognition system is briefly introduced in Section 2. Section 3 detailedly describes emotional feature extraction, including acoustic feature extraction and facial feature extraction. Multimodal emotional database used for experiments is presented in Section 4. Section 5 describes the used emotion classifier named support vector machines (SVM). Section 6 shows the experiment results and analysis. Finally, the conclusions are given in Section 7.

2 Multimodal Emotion Recognition System

As shown in Fig.1, a fundamental multimodal emotion recognition system is comprised of two steps: emotional feature extraction, emotion classification. Emotional feature extraction aims at deriving suitable features from facial image samples or affective speech samples, which can efficiently characterize different emotions of human expression. In detail, for multimodal emotion recognition, emotional feature extraction contains acoustic feature extraction from affective speech samples, and facial feature extraction from facial image samples. Emotion classification is concerned with identifying different emotions by using a suitable emotion classifier. It's worth pointing out that on multimodal emotion recognition tasks it's important to employ suitable multimodal modality fusion techniques to perform multimodal emotion recognition task. So far, two typical multimodal modality fusion techniques contain the feature-level fusion method and the decision-level fusion method.

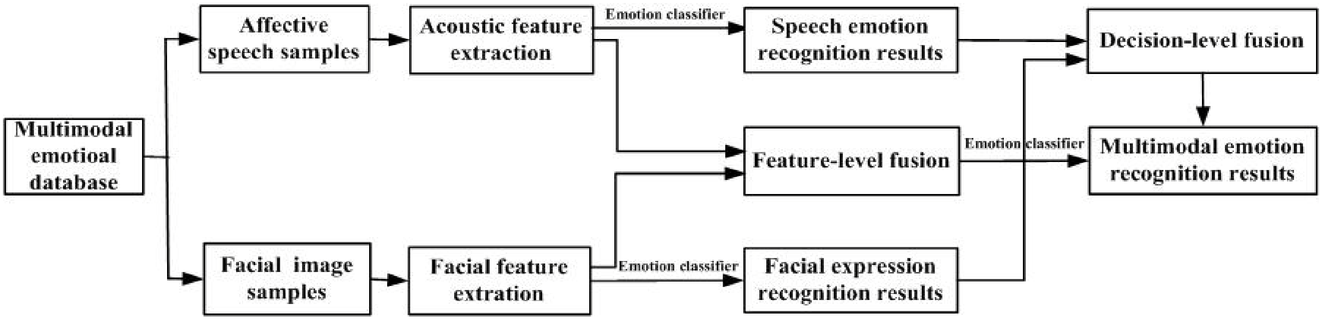
Multimodal emotioal database
Affective speech samples
Acoustic feature extraction
Emotion classifier
Speech emotion recognition results
Decision-level fusion
Feature-level fusion
Emotion classifier
Multimodal emotion recognition results
Facial image samples
Facial feature extration
Emotion classifier
Facial expression recognition results

Fig.1 A flowchart of a multimodal emotion recognition system

Generally, modality fusion techniques focus on integrating all single modalities into a combined representation. Fusion of information from multiple sources is usually performed at the feature-level or the decision-level. To ensure that the fusion issue is tractable, the individual modalities are usually assumed to be independent for each other.

Feature-level fusion is performed by concatenating the extracted features from each modality into a larger feature vector. The resulting feature vector is then input into a single classifier to perform classification tasks. Decision-level fusion enables each modality to be firstly pre-classified independently and the final classification results are decided by the fusion of the outputs of the different modalities. The quality of decision-level fusion depends on the chosen framework for optimality. At present, designing optimal strategies for decision-level fusion is still an open research issue. The widely used rules for decision-level fusion contain the maximum/minimum rule, the sum rule, the product rule, the average rule, majority vote, etc. Each fusion method has its two sides. Feature level fusion benefits of interdependence and correlation of the affective features in both modalities, but is criticized for ignoring the differences in temporal structure, scale and metrics. By contrast, decision level fusion attempts are criticized for ignoring the mentioned correlation and complementary role of the modalities. In this work, we will investigate the performance of fusion methods at the feature-level and the decision-level on the multimodal emotion recognition tasks.

3 Emotional Feature Extraction

In this section, we will present the details of emotional feature extraction, including acoustic feature extraction as well as facial feature extraction.

3.1 Acoustic Feature Extraction

It's well-known that there exists a large number of paralinguistic and linguistic feature information characterizing human emotion expression in affective speech signals. However, so far there are still many debates about the fact which are the best features for speech emotion recognition. Nevertheless, in many literatures about speech emotion recognition, three types of acoustic features, including prosody features, voice quality features as well as the spectral features, are the most popular features for speech emotion recognition. In this work, we extract the typical prosody features, such as pitch, intensity and duration. And the extracted voice quality features are consisted of the first three formants (F1, F2, F3), spectral energy distribution, harmonics-to-noise-ratio (HNR), pitch irregularity (jitter) and amplitude irregularity (shimmer). As one of representative spectral features, the well-known Mel-frequency Cepstral Coefficients (MFCC) features are extracted. For each acoustic feature, some typical statistical parameters, such as mean, median, quartiles and standard derivations, are computed. The software used to extract all the acoustic features extraction is PRAAT, which is a shareware program designed by Paul Boersma and David Weenink of the Institute of Phonetics Sciences of the University of Amsterdam and publicly available online at http://www.praat.org.

3.1.1 Prosody features

Prosody involves in the stress and intonation patterns of spoken language [8]. Intuitively, prosody is usually referred to as the first acoustic parameter when considering acoustic feature extraction due to its importance in conveying emotional expression. The widely used prosody features extracted in this work contain pitch, intensity, and duration, as described as follows:

◆ Pitch-related parameters:

Pitch, also regarded as fundamental frequency (F0), is used to estimate the rate of vocal fold vibration and is taken as one of the most important feature attributes in human emotion expression and detection from affective speech signals [9, 10].

It has been proved that the pitch contour varies with the emotional states being expressed. For instance, due to the sex difference, anger and joy shows a relatively higher pitch mean values, whereas the mean values of boredom and sadness are slightly slower compared to the neutral emotion [11]. In this work, we employed a robust version of the autocorrelation algorithm, default used in the PRAAT, to automatically compute the pitch contour for each utterance [12]. Finally, based on the pitch contour of each utterance, ten statistics are extracted: mean, maximum, minimum, range, standard deviation, first quartile, third quartile, median, inter-quartile range, the mean-absolute-slope

◆ Intensity -related parameters:

Intensity, denoted by the volume or energy of the speech, is correlated with loudness and is one of the most intuitive indicators in the relation voice-emotion [13, 14]. Even if we are not experts in this matter, we could easier imagine someone angry shouting than gently whispering. The intensity contour provides information that can be used to differentiate sets of emotions. Higher intensity levels are found in those with high arousal levels such as anger, surprise, and joy, while sadness and boredom with low arousal levels yield lower intensity values [11, 15]. The algorithm used to calculate the intensity convolutes a Kaiser-20 window over the speech signal–the default procedure in PRAAT. From global statistics directly derived from the intensity contour, we selected 9 statistics: maximum, minimum, range, mean, standard deviation, first quartile, median, third quartile, inter-quartile range.

◆ Duration related parameters:

Prosody involves also duration-related measurements. One of the most important durational measurements in the aim to discriminate among speaker's emotional states is the speaking rate. An acoustic correlated of the speaking rate can be defined as the inverse of the average of the voiced region length within a certain time interval. It has been noted that fear, disgust, anger, and joy often have an increased speaking rate, while sadness has a reduced articulation rate with irregular pauses [15]. Since we do not know the onset time and duration of the individual phonemes, measures of the speaking rate are obtained with respect to voiced and unvoiced regions. Statistics are calculated from individual durations of voiced and unvoiced regions, which are extracted from the pitch contour. All measures are calculated with respect to the length of the utterance. The duration of each frame is 20ms. For the duration related parameters, we selected 6 statistics: total-frames, voiced-frames, unvoiced-frames, ratio of voiced vs. unvoiced frames, ratio of voiced-frames vs. total-frames, ratio of unvoiced-frames vs. total-frames.

3.1.2 Voice quality features

Voice quality is referred to as the characteristic auditory colouring of an individual's voice, derived from a variety of laryngeal and supralaryngeal features and running continuously through the individual's speech [16]. A wide range of phonetic variables contribute to the subjective impression of voice quality. Voice quality is usually changed to strengthen the impression of emotions. Voice quality measures, which have been directly related to emotions, include the first three formants, spectral energy distribution, harmonics-to-noise ratio (HNR), pitch irregularity (jitter) and amplitude irregularity (shimmer) [17-19].

◆Formants related parameters:

The resonant frequencies produced in the vocal tract are referred to as formant frequencies or formants [20]. Each formant is characterized by its center frequency and its bandwidth. It has been found that the first three formants (F1, F2, F3) are affected by the emotional states of speech more than the other formants [21]. It was also noticed that the amplitudes of F2 and F3 were higher with respect to that of F1 for anger and fear compared with neutral speech [22]. To estimate the formants, PRAAT applies a Gaussian-like window for each analysis window and computes the linear predictive coding (LPC) coefficients with the Burg algorithm [23], which is a recursive estimator for auto-regressive models, where each step is estimated using the results from the previous step. The following statistics are measured for the extracted formant parameters: mean of F1, std of F1, median of F1, bandwidth of median of F1, mean of F2, std of F2, median of F2, bandwidth of median of F2, mean of F3, std of F3, median of F3, bandwidth of median of F3.

◆Spectral energy distribution related parameters:

The spectral energy distribution is calculated within four different frequency bands in order to decide, whether the band contains mainly harmonics of the fundamental frequency or turbulent noise [19]. There are many contradictions in identifying the best frequency band of the power spectrum in order to classify emotions.

Many investigators put high significance on the low frequency bands, such as the 0-1.5 kHz band [21, 24], whereas other suggest the opposite [15]. Here, spectral energy distribution in four different frequency bands including low and high frequency (0-5kHz) bands are calculated directly by PRAAT. The following features are measured: band energy from 0 Hz to 500 Hz, band energy from 500 Hz to 1000 Hz, band energy from 2500 Hz to 4000 Hz, band energy from 4000 Hz to 5000 Hz.

◆ Harmonics-to-noise-ratio related parameters:

The harmonic-to-noise ratio (HNR) is defined as the ratio of the energy of the harmonic part to the energy of the remaining part of the signal and represents the degree of acoustic periodicity. HNR estimation can be considered as an acoustic correlation with breathiness and roughness [25]. The values of HNR in the sentences expressed with anger are significantly higher than the neutral expression [25].

The algorithm performs acoustic periodicity detection on the basis of an accurate autocorrelation method [12]. The following features are measured: maximum, minimum, range, mean, standard deviation.

◆Jitter and Shimmer:

Jitter/shimmer measures have been considered in voice quality assessment to describe the kinds of irregularities associated with vocal pathology [26].

Jitter: It measures the cycle-to-cycle variations of the fundamental period averaging the magnitude difference of consecutive fundamental periods, divided by the mean period.

Jitter is defined as the relative mean absolute third-order difference of the point process, which is exceptionally calculated using PRAAT [27]. Jitter is calculated with the following Eq. (1), in which T_i is the i -th peak-to-peak interval and N is the number of intervals:

$$Jitter(\%) = \sum_{i=2}^{N-1}(2T_i - T_{i-1} - T_{i+1}) \Big/ \sum_{i=2}^{N-1} T_i \qquad (1)$$

Shimmer: It measures the cycle-to-cycle variations of amplitude by averaging the magnitude difference of the amplitudes of consecutive periods, divided by the mean amplitude [26].

Shimmer is calculated similarly to Jitter as shown in Eq. (2), in which E_i is the i -th peak-to-peak energy values and N is the number of intervals:

$$Shimmer(\%) = \sum_{i=2}^{N-1}(2E_i - E_{i-1} - E_{i+1}) \Big/ \sum_{i=2}^{N-1} E_i \quad (2)$$

3.1.3 MFCC features

As the representative spectral features, for each utterance the first 13 Mel-frequency cepstral coefficients (MFCC) (including log-energy) with their first-delta and second-delta components are extracted by using a 25 ms Hamming window at intervals of 10 ms. The mean and std of MFCC as well as its first-delta and second-delta components is computed for each utterance, giving 156 MFCC features.

To summarize, for each utterance from the emotional speech corpus, 25 prosody features, 23 voice quality features as well as 156 MFCC features are extracted. These extracted 204 features in total are statistical in Table 1.

Table 1 Acoustic feature extraction

Feature types	Feature groups	Statistics
Prosody features	Pitch	maximum, minimum, range, mean, std, first quartile, median, third quartile, inter-quartile range, the mean-absolute-slope
	Intensity	maximum, minimum, range, mean, std, first quartile, median, third quartile, inter-quartile range
	Duration	total-frames, voiced-frames, unvoiced-frames, ratio of voiced vs. unvoiced frames, ratio of voiced-frames vs. total-frames, ratio of unvoiced-frames vs. total-frames
Voice quality features	Formants	mean of F1, std of F1, median of F1, bandwidth of median of F1, mean of F2, std of F2, median of F2, bandwidth of median of F2, mean of F3, std of F3, median of F3, bandwidth of median of F3
	Spectral energy distribution	band energy from 0 Hz to 500 Hz, band energy from 500 Hz to 1000 Hz, band energy from 2500 Hz to 4000 Hz, band energy from 4000 Hz to 5000 Hz.
	HNR	maximum, minimum, range, mean, std
	Jitter, Shimmer	Jitter, Shimmer
Spectral features	MFCC	mean, std of the first 13 MFCC, and their first-deltas and second-deltas

3.2 Facial Feature Extraction

Facial feature extraction is to extract facial features to represent the facial changes caused by facial expressions. Two types of features, i.e., geometric features and appearance features, are usually used for facial representation [28]. Geometric features present the shape and locations of facial components such as mouth, eyes, brows, and nose. The facial components or facial feature points are extracted to form a feature vector that represents the face geometry. Fiducial facial feature points have been widely adopted as geometric features for facial representation. For instance, the geometric positions of 34 fiducial points on a face are usually used to represent facial images [29]. In contrast to geometric features, appearance features encode changes in skin texture such as wrinkles, bulges and furrows. The representative appearance features contains the raw pixels of facial images, Gabor wavelets representation [30], Eigenfaces [31], and Fisherfaces [32], etc. In recent years, a new face descriptor called local binary patterns (LBP) [33], originally proposed for texture analysis and a non-parametric method efficiently summarizing the local structures of an image, have received increasing interest for facial expression representation. The most important property of LBP features is their tolerance against illumination changes and their computational simplicity. LBP has been successfully applied as an appearance feature extraction for facial expression recognition [34-36]. Especially, in recent years, in our previously published work [37-38] we have successfully adopted LBP as facial feature representation for facial expression recognition and reported promising performance. Therefore, in this work we will extract the LBP features from facial images for facial expression recognition.

The LBP features extraction is introduced as follows.

The original local binary patterns (LBP) [33] operator takes a local neighborhood around each pixel, thresholds the pixels of the neighborhood at the value of the central pixel and uses the resulting binary-valued image patch as a local image descriptor. It was originally defined for 3 × 3 neighborhood, giving 8 bit codes based on the 8 pixels around the central one. The operator labels the pixels of an image by thresholding a 3 × 3 neighborhood of each pixel with the center value and considering the results as a binary number, and the 256-bin histogram of the LBP labels computed over a region is used as a texture descriptor. Figure 1 gives an example of the basic LBP operator.

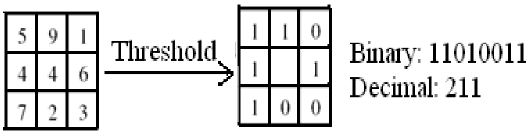

Fig 1. An example of the basic LBP operator

The process of LBP features extraction generally is consisted of three steps: firstly, a facial image is divided into several non-overlapping blocks. Secondly, LBP histograms are computed for each block. Finally, the block LBP histograms are concatenated into a single vector. As a result, the facial image is represented by the LBP code.

For the LBP features extraction, the pre-processing procedure of facial images is summarized as follows. As done in [35, 37], we normalized the faces to a fixed distance of 55 pixels between the two eyes. Automatic face registration can be achieved by a robust real-time face detector developed by Viola and Jones [39]. From the results of automatic face detection, such as face location, face width and face height, two square bounding boxes for left eye and right eye are created respectively. Then, two eyes location can be quickly worked out in terms of the centers of two square bounding boxes for left eye and right eye. Based on the two eyes location, facial images of 110×150 pixels were cropped from original frames. No further alignment of facial features such as alignment of mouth was performed in our work.

The cropped facial images of 110 × 150 pixels contain facial main components such as mouth, eyes, brows and noses. The LBP operator is applied to the whole region of the cropped facial images. For better uniform-LBP feature extraction, two parameters, i.e., the LBP operator and the number of regions divided, need to be optimized. Similar to the setting in [35, 37], we selected the 59-bin operator, and divided the 110 × 150 pixels face images into 18 × 21 pixels regions, giving a good trade-off between recognition performance and feature vector length. Thus face images were divided into 42 (6 × 7) regions, and represented by the LBP histograms with the length of 2478 (59 × 42).

4 Multimodal Emotional Database

To evaluate the performance of multimodal emotion recognition, the publicly available eNTERFACE’05 [7] multimodal emotion database is used for experiments. It contains six basic emotions, i.e., anger, disgust, fear, joy, sadness, and surprise. 43 non-native (eight female) English speaking subjects from 14 nations posed the six basic emotions with five sentences. Each subject was told to listen to six successive short stories, each of them intended to elicit a particular emotion. They then had to react to each of the situations by uttering previously read phrases that fit the short story. Five phrases are available per emotion, as “I have nothing to give you! Please don’t hurt me!” in the case of fear. Two experts judged whether the reaction expressed the intended emotion in an unambiguous way. Only if this was the case was a sample (= sentence) added to the database. Therefore, each sentence in the database has one assigned emotion label, which indicates the emotion expressed by the speaker in this sentence.

In this work, 30 samples (5 samples per emotion) for each subject are selected. Finally, 1290 samples in total are used for experiments. Each of facial images in this database has a resolution of 720 × 576 pixels. The cropped facial images of 110 × 150 pixels, which contain facial main components such as mouth, eyes, brows and noses, are shown in Figure 2.

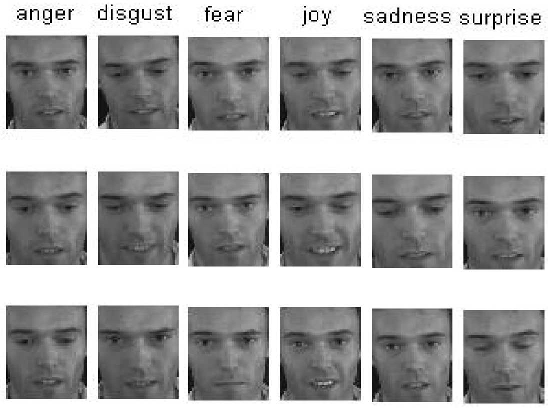

Fig 2. Sample images from the eNTERFACE’05 database

5 Emotion Classifier

Emotion classification maps feature vectors onto emotion classes through a classifier’s learning by data examples. After feature extraction, the accuracy of emotion recognition relies heavily on the use of a good pattern classifier. At present, the representative emotion classifier, such as linear discriminant classifiers (LDC), K-nearest-neighbor (KNN), artificial neural network (ANN), and support vector machines (SVM), have been successfully applied for emotion recognition. Among them, SVM has become the most popular classification tool due to its strong discriminating capability. In this work, we will use SVM for emotion recognition. The basic principle of SVM is given below.

Support vector machines (SVM) is a relatively new machine learning algorithm developed by Vapnik [40]. Based on the statistical learning theory of structural risk management, SVM aims to transform the input vectors to a higher dimensional space by a nonlinear transform, and then an optimal hyperplane which separates the data can be found.

Given the training data set $(x_1, y_1), ..., (x_l, y_l), y_i \in \{-1,1\}$, to find the optimal hyperplane, a nonlinear transform, $Z = \Phi(x)$, is used to make training data become linearly dividable. A weight w and offset b satisfying the following criteria will be found:

$$\begin{cases} w^T z_i + b \geq 1, & y_i = 1 \\ w^T z_i + b \leq -1, & y_i = -1 \end{cases} \tag{3}$$

The above procedure can be summarized to the following:

$$\min_{w,b} \Phi(w) = \frac{1}{2}(w^T w) \tag{4}$$

subject to $y_i(w^T z_i + b) \geq 1, \quad i = 1, 2, ..., n$

If the sample data is not linearly dividable, the following function should be minimized.

$$\Phi(w) = \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{5}$$

whereas ξ can be understood as the error of the classification and C is the penalty parameter for this term.

By using the Lagrange method, the decision function of $w_0 = \sum_{i=1}^{l} \lambda_i y_i z_i$ will be

$$f = \mathrm{sgn}[\sum_{i=0}^{l} \lambda_i y_i (z^T z_i) + b] \tag{6}$$

From the functional theory, a non-negative symmetrical function $K(u,v)$ uniquely defines a Hilbert space H, where K is the rebuild kernel in the space H:

$$K(u,v) = \sum_i \alpha \varphi_i(u) \varphi_i(v) \tag{7}$$

This stands for an internal product of a characteristic space:

$$z_i^T z = \Phi(x_i)^T \Phi(x) = K(x_i, x) \tag{8}$$

Then the decision function can be written as:

$$f = \mathrm{sgn}[\sum_{i=1}^{l} \lambda_i y_i K(x_i, x) + b] \tag{9}$$

The development of a SVM classification model depends on the selection of kernel function. There are four typical kernels that can be used in SVM models. These include linear, polynomial, radial basis function (RBF) and sigmoid function, as described below.

The linear kernel function is defined as

$$K(x_i, x_j) = x_i^T x_j \tag{10}$$

The polynomial kernel function is defined as

$$K(x_i, x_j) = (\gamma x_i^T x_j + coefficient)^{degree} \tag{11}$$

The RBF kernel function is defined as

$$K(x_i, x_j) = \exp(-\gamma \, | x_i - x_j |^2 \tag{12}$$

The sigmoid kernel function is defined as

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + coefficient) \tag{13}$$

Many real-world data sets involve multi-class problem. Since SVM is inherently binary classifiers, the binary SVM is needed to extend to be multi-class SVM for multi-class problem. Currently, there are two types of approaches for building multi-class SVM. One is the “single machine” approach, which attempts to construct multi-class SVM by solving a single optimization problem. The other is the “divide and conquer” approach, which decomposes the multi-class problem into several binary sub-problems, and builds a standard SVM for each. The most popular decomposing strategy is probably the “one-against-all”. The “one-against-all” approach consists of building one SVM per class and aims to distinguish the samples in a single class from the samples in all remaining classes. Another popular decomposing strategy is the “one-against-one”. The “one-against-one” approach builds one SVM for each pair of classes. When applied to a test point, each classification gives one vote to the winning class and the point is labeled with the class having most votes.

6 Experiments

In this section, we perform facial expression recognition experiments on the benchmarking eNTERFACE’05 database, and present experimental results and analysis.

In all classification experiments, the SVM classifier is used. We used the LIBSVM package [41] to implement the SVM algorithm with radial basis function (RBF) kernel, kernel parameter optimization, one-against-one strategy for multi-class classification problem. With regard to the parameter optimization of SVM, we carried out grid-search on the hyper-parameters in the 10-fold cross-validation on the training sets. All extracted features, including the LBP features and the acoustic features, were normalized by a mapping to [0, 1] before anything else. As done in [34-37], a 10-fold cross validation scheme is employed in 6-class emotion classification experiments, and the average results are reported.

6.1 Experiments with Single Emotion Recognition

In this section, we will present speech emotion recognition results and facial expression recognition results, respectively.

6.1.1 Experiments with Speech Emotion Recognition

We used the extracted 204 acoustic features to perform speech emotion recognition. Table 1 presents the confusion matrix of recognition results with SVM. As shown in Table 1, it can be seen that “anger” and “sadness” could be discriminated well with an accuracy of 77.67%, and 74.88%, respectively, while other four emotions could be classified with relatively lower accuracies. In detail, the recognition accuracy is 62.79% for surprise, 58.6% for joy, 52.56% for fear, and 50.23% for disgust. The overall accuracy for speech emotion recognition is 62.79%.

6.1.2 Experiments with Facial Expression Recognition

The extracted LBP features with the length of 2478 are directly used to classify facial expression. The confusion matrix of recognition results with SVM classifier is given in Table 2. The confusion matrix in Table 2 indicates that “joy” could be discriminated best with an accuracy of 60%, while other five emotions could be classified with a low accuracy of less than 50%. The overall accuracy for facial expression recognition is 44.73%. This shows that the facial images from the eNTERFACE’05 database have low quality recordings and need to be improved further. Compared with the results in Table 1, Table 2 shows that speech is more effective than face on the emotion recognition tasks.

6.2 Experiments with Multimodal Emotion Recognition

We firstly fused the extracted LBP features and the acoustic features and then used SVM to perform multimodal emotion recognition experiments at the feature-level. The confusion matrix of recognition results with SVM at the feature-level is presented in Table 3. From Table 3, we can observe that three emotions, i.e., anger, joy and sadness, can be recognized well. In detail, the recognition accuracy is 81.86% for anger, 73.02% for joy, and 74.88% for sadness. The other four emotions are classified badly with an accuracy of less than 61%. The overall recognition performance of multimodal emotion recognition at the feature-level is 66.51%, making an about 4% improvement over speech emotion recognition performance, and about 18% improvement over facial expression recognition performance, respectively. This indicates that fusion of facial expression and speech at the feature-level achieves better performance than the used mono-modality (i.e., facial expression or speech).

We subsequently performed multimodal emotion recognition experiments at the decision-level. The typical six rules, including the maximum/minimum rule, the sum rule, the product rule, the average rule, and majority vote, were used for decision-level fusion. The recognition performance of different used fusion rules at the decision-level is given in Table 4. We can see that in Table 4 the "product" rule gives the best accuracy of 67.44%, outperforming the other five fusion rules. In addition, the performance obtained by the "product" rule at the decision-level is also higher than the performance at the feature-level. The confusion matrix of the "product" rule obtained by the "product" rule is presented in Table 5.

Table 1. Confusion matrix of speech emotion recognition results

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Accuracy
Anger	167	11	12	7	3	15	77.67%
Disgust	27	108	16	35	17	12	50.23%
Fear	20	15	113	11	35	21	52.56%
Joy	17	36	7	126	2	27	58.60%
Sadness	4	10	21	6	161	13	74.88%
Surprise	8	17	19	23	13	135	62.79%
Overall accuracy							62.79%

Table 2. Confusion matrix of facial expression recognition results

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Accuracy
Anger	106	25	20	18	13	33	49.30%
Disgust	42	89	26	34	10	14	41.39%
Fear	35	27	70	18	43	22	32.56%
Joy	20	18	13	129	12	23	60.00%
Sadness	18	16	44	13	98	26	45.58%
Surprise	42	14	18	28	28	85	39.53%
Overall accuracy							44.73%

Table 3 Confusion matrix of multimodal emotion recognition results at the feature-level

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Accuracy
Anger	176	13	9	4	2	11	81.86%
Disgust	28	112	24	25	10	16	52.09%
Fear	23	21	122	6	27	16	56.74%
Joy	7	25	5	157	4	17	73.02%
Sadness	4	11	26	2	161	11	74.88%
Surprise	12	17	26	16	14	130	60.46%
Overall accuracy							66.51%

Table 4 Multimodal emotion recognition results at the decision-level with different fusion rules

Rules	Maximum	Minimum	Sum	Product	Average	Majority vote
Accuracy	63.72%	65.35%	65.51%	67.44%	66.51%	53.41%

Table 5 Confusion matrix of multimodal emotion recognition results at the decision-level with the “product” rule

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Accuracy
Anger	169	11	12	8	3	12	78.60%
Disgust	22	113	21	35	13	11	52.56%
Fear	18	24	120	5	28	20	55.81%
Joy	5	22	6	162	2	18	75.35%
Sadness	3	10	27	3	159	13	73.95%
Surprise	9	9	20	16	14	147	68.37%
Overall accuracy							67.44%

7 Conclusions

This paper has presented a multimodal emotion recognition method integrating facial expression and affective speech. We separately performed facial expression recognition and speech emotion recognition experiments, and then fused facial expression and affective speech modalities for multimodal emotion recognition at the feature-level and at the decision-level. From the experimental results, we can conclude that the multimodal emotion recognition method fusing facial expression and affective speech, obtains better recognition accuracy in general, outperforming the classification performance using the single facial expression modality or the single speech modality. Additionally, among the used six rules for decision-level fusion, the product rule presents the highest accuracy of 67.44%, outperforming the obtained performance at the feature-level fusion.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant No.61203257 and No.61272261, and the Scientific Research Fund of Hunan Provincial Education Department under Grant No.10C0578.

References:

[1] R Picard, "Affective computing," Cambridge, USA: MIT Press, 1997.

[2] R Cowie, E Douglas-Cowie, N Tsapatsoulis *et al.*, Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine*, Vol. 18, No. 1, 2001, pp. 32-80.

[3] N Fragopanagos, and J G Taylor, Emotion recognition in human-computer interaction, *Neural Networks*, Vol. 18, No. 4, 2005, pp. 389-405.

[4] S Ramakrishnan, and I M M El Emary, Speech emotion recognition approaches in human computer interaction, *Telecommunication Systems*, 2011, pp. 1-12.

[5] S Zhang, X Zhao, and B Lei, Speech Emotion Recognition Using an Enhanced Kernel Isomap for Human-Robot Interaction, *International Journal of Advanced Robotic Systems*, Vol. 10, 2013, pp. 1-7.

[6] G Kharat, and S Dudul, Human emotion recognition system using optimally designed SVM with different facial feature extraction techniques, *WSEAS Transactions on Computers*, Vol. 7, No. 6, 2008, pp. 650-659.

[7] O Martin, I Kotsia, B Macq *et al.*, The eNTERFACE'05 Audio-Visual Emotion Database, Proc. 22nd International Conference on Data Engineering Workshops 2006, pp.

[8] A Batliner, A Buckow, H Niemann *et al.*, The prosody module, *VERBMOBIL: Foundations of Speech-to-speech Translations*, 2000, pp. 106–121.

[9] J Ang, R Dhillon, A Krupski *et al.*, Prosody-based automatic detection of annoyance and frustration in human-computer dialog, Proc. 7th International Conference on Spoken Language Processing (ICSLP'02), 2002, pp. 2037-2040.

[10] V Petrushin, Emotion recognition in speech signal: experimental study, development, and application, Proc. 6th International Conference on Spoken Language Processing (ICSLP'00), 2000, pp. 222-225.

[11] I Murray, and J Arnott, Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *The Journal of the Acoustical Society of America*, Vol. 93, 1993, pp. 1097-1108.

[12] P Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, *Proceedings of the Institute of Phonetic Sciences*, Vol. 17, 1993, pp. 97-110.

[13] S McGilloway, R Cowie, E Douglas-Cowie *et al.*, Approaching automatic recognition of emotion from voice: a rough benchmark, Proc. the ISCA Workshop on Speech and Emotion, 2000, pp. 207-212.

[14] T Polzin, and A Waibel, Emotion-sensitive human-computer interfaces, Proc. the ISCA Workshop on Speech and Emotion, 2000, pp. 201-206.

[15] T L Nwe, S W Foo, and L C De Silva, Speech emotion recognition using hidden Markov models, *Speech Communication*, Vol. 41, No. 4, 2003, pp. 603-623.

[16] R Trask, *A dictionary of phonetics and phonology*, Routledge, London: Burns & Oates, 1996.

[17] G Klasmeyer, and W Sendlmeier, Voice and emotional states, *Voice Quality Measurement*, 2000, pp. 339-358.

[18] G Klasmeyer, The perceptual importance of selected voice quality parameters, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97), 1997, pp. 1615-1618.

[19] G Klasmeyer, and W Sendlmeier, Objective voice parameters to characterize the

emotional content in speech, Proc. 13th International Congress Phonetic Sciences (ICPhS’95), 1995, pp. 182-185.

[20] L Rabiner, and R Schafer, *Digital processing of speech signals*, New Jersey: Prentice-hall Englewood Cliffs, 1978.

[21] F Tolkmitt, and K Scherer, Effect of experimentally induced stress on vocal parameters, *Journal of Experimental Psychology: Human Perception and Performance,* Vol. 12, No. 3, 1986, pp. 302-313.

[22] C Williams, and K Stevens, Emotions and speech: Some acoustical correlates, *Journal of the Acoustical Society of America,* Vol. 52, No. 4B, 1972, pp. 1238-1250.

[23] J Pittam, and K Scherer, "Vocal expression and communication of emotion," *In M. Lewis & J. M. Haviland (Eds.), Handbook of emotions*, pp. 185-197, New York: Guilford Press, 1993.

[24] R Banse, and K R Scherer, Acoustic profiles in vocal emotion expression, *Journal of Personality and Social Psychology,* Vol. 70, 1996, pp. 614-636.

[25] K Alter, E Rank, S Kotz *et al.,* Accentuation and emotions-two different systems?, Proc. ITRW on Speech and Emotion, 2000, pp. 138-142.

[26] D Michaelis, M Fr hlich, and H Strube, Selection and combination of acoustic features for the description of pathologic voices, *Journal of the Acoustical Society of America,* Vol. 103, No. 3, 1998, pp. 1628-1639.

[27] H Kasuya, Y Endo, and S Saliu, Novel acoustic measurements of jitter and shimmer characteristics from pathological voice, Proc. EUROSPEECH '93, 1993, pp. 1973-1976.

[28] Y Tian, T Kanade, and J F Cohn, Facial Expression Recognition, *Handbook of face recognition*, 2011, pp. 487-519.

[29] M Kyperountas, A Tefas, and I Pitas, Salient feature and reliable classifier selection for facial expression classification, *Pattern Recognition,* Vol. 43, No. 3, 2010, pp. 972-986.

[30] W Gu, C Xiang, Y Venkatesh *et al.,* Facial expression recognition using radial encoding of local Gabor features and classifier synthesis, *Pattern Recognition,* Vol. 45, No. 1, 2012, pp. 80-91.

[31] M A Turk, and A P Pentland, Face recognition using eigenfaces, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1991, pp. 586-591.

[32] P N Belhumeur, J P Hespanha, and D J Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 19, No. 7, 1997, pp. 711-720.

[33] T Ojala, M Pietik inen, and T M enp, Multiresolution gray scale and rotation invariant texture analysis with local binary patterns, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 24, No. 7, 2002, pp. 971-987.

[34] C Shan, S Gong, and P McOwan, Robust facial expression recognition using local binary patterns, Proc. IEEE International Conference on Image Processing (ICIP), 2005, pp. 370-373.

[35] C Shan, S Gong, and P McOwan, Facial expression recognition based on Local Binary Patterns: A comprehensive study, *Image and Vision Computing,* Vol. 27, No. 6, 2009, pp. 803-816.

[36] S Moore, and R Bowden, Local binary patterns for multi-view facial expression recognition, *Computer Vision and Image Understanding,* Vol. 115, No. 4, 2011, pp. 541-558.

[37] S Zhang, X Zhao, and B Lei, Facial Expression Recognition Based on Local Binary Patterns and Local Fisher Discriminant Analysis, *WSEAS TRANSACTIONS on Signal Processing,* Vol. 8, No. 1, 2012, pp. 21-31.

[38] S Zhang, X Zhao, and B Lei, Facial Expression Recognition Using Sparse Representation, *WSEAS Transaction on Systems,* Vol. 11, No. 8, 2012, pp. 440-452.

[39] P Viola, and M Jones, Robust real-time face detection, *International Journal of Computer Vision,* Vol. 57, No. 2, 2004, pp. 137-154.

[40] V Vapnik, *The nature of statistical learning theory*: Springer-Verlag, New.York, 2000.

[41] C Chang, and C Lin, LIBSVM: a library for support vector machines, 2001, *Software available at http://www. csie. ntu. edu. tw/cjlin/libsvm*, 2001.