RECOGNIZING HUMAN EMOTION FROM AUDIOVISUAL INFORMAITON

Yongjin Wang and Ling Guan

Department of Electrical and Computer Engineering, Ryerson University
Toronto, Ontario, Canada M5B 2K3
{ywang, lguan}@ee.ryerson.ca

ABSTRACT

In this paper, we present an emotion recognition system to classify human emotional state from audiovisual signals. We extract prosodic, Mel-Frequency Cepstral Coefficient (MFCC), and formant frequency features to represent the audio characteristics of the emotional speech. A face detection scheme based on HSV color model is used to detect the face from the background. The facial expressions are represented by Gabor wavelet features. We perform feature selection by using a stepwise method based on Mahalanobis distance. A classification scheme involving the analysis of individual class and combinations of different classes is proposed. Our emotion recognition system is tested over a language and race independent database, and an overall recognition accuracy of 82.14% is achieved.

1. INTRODUCTION

Recognizing human emotion by computer has been an active research area in the past a few years. An efficient human emotion recognition system will help to make the interaction between human and computer more natural and friendly. It has broad applications in areas such as education, entertainment, customer service, etc. As two of the major indicators of human affective state, speech and facial expression play important roles in emotion recognition. In this paper, we present an emotion recognition system to recognize human emotional state from these two modalities.

A wide investigation on the dimensions of emotions has been performed in the past. One popular way of representing emotional state is to categorize them into six principal emotions: *happiness (HA), sadness (SA), anger (AN), fear (FE), surprise (SU), and disgust (DI)*. The other emotions can be regarded as the combination or variation of these six principal emotions. These six emotions are the focus of our study in this paper.

The majority of the recent work in the field either focused on speech alone [1.2], or facial expression only [3]. However, as shown in [4], some of the emotions might be audio dominant, while the others are visual dominant. The combination of audio and visual data will convey more information about the human emotional state. The complementary relationship of these two modalities on different emotions will help to achieve higher recognition accuracy.

Another problem for the research of human emotion recognition is that there is no standard database available. Most of the systems are restricted to a database of only one language, and one race. However, the way people convey their emotional state in speech might be different according to their cultural background and language, and the facial expression might be also effected by racial aspects such as skin color and facial hair. An efficient emotion recognition system must be able to adapt itself to these aspects.

In this paper, we study the audiovisual recognition of human emotion regardless of the subject's cultural background, language, and race. The audio features including prosodic, MFCC, and formant frequency features are extracted from the speech to map the emotional speech to the corresponding feature space. A group of Gabor wavelet features are extracted to represent the facial expressions. The stepwise method, which involves maximizing the between-class Mahalanobis distance, is used to select the significant features. We proposed a multi-classifier scheme based on Fisher's Linear Discriminant Analysis (FLDA) to classify emotional video samples.

The remainder of this paper is organized as follows. Section 2 describes the database that we used to test the proposed system. The methods that we use for emotion recognition are described in section 3. Section 4 presents the experimental results. Discussions and conclusions are given in section 5.

2. DATA ACQUISITION

In order to conduct the research of recognizing human emotion, a video database that can truly convey the emotional state of human was first collected. For a more general application, the data should not be restricted to the user's cultural background, language, race, etc. To ensure the diversity of the database, we collected video samples from eight subjects, speaking six different languages. The six languages are English, Chinese (Mandarin), Urdu, Punjabi, Persian, and Italian. Different accents of English and Chinese were also included. For each emotional class, more than ten reference sentences were provided to the subjects. Some of the subjects have facial hair, which also further increase the diversity of the database. We collected a total of 500 video samples. The samples were recorded at a sampling rate of 22050 Hz, using a single channel 16-bit digitization.

3. EMOTION RECOGNITION SYSTEM

As shown in Figure 1, the proposed human emotion recognition system is composed of four components: audio feature extraction, visual feature extraction, feature selection, and classification.

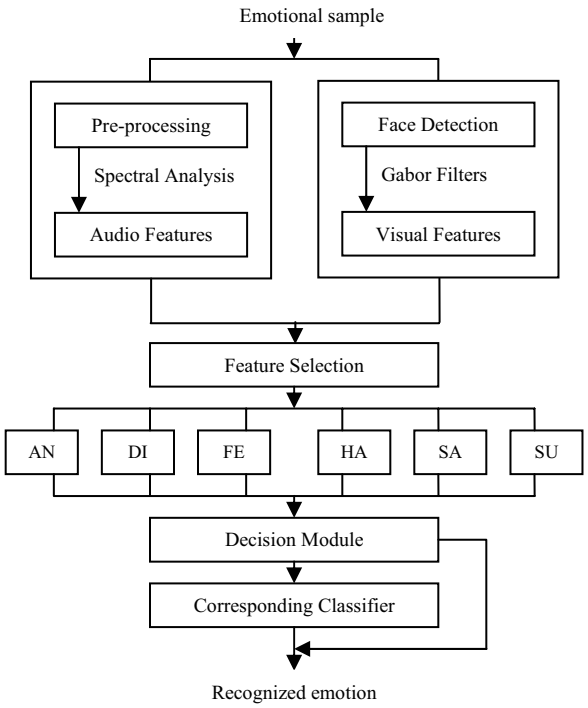

Figure 1: Proposed Emotion Recognition System

3.1 Audio Feature Extraction

3.1.1 Pre-processing

To reduce the effects of noise in the speech utterance, we perform noise reduction at the preprocessing stage. The "hiss" of the recording machine and background noise are reduced by thresholding the wavelet coefficients. Compared with the traditional low pass filtering method, the wavelet method has the advantage of reducing the noise efficiently without blurring the features in the original signal [5]. In order to exclude the silence periods which do not contribute to human emotion, we perform leading and trailing edge elimination on the noise-reduced signal. We estimate the maximum noise amplitude in a predefined short period of time. The threshold is calculated by adding a noise margin to the maximum amplitude. This noise margin value is estimated based on experiments. We then exclude the leading and trailing silence by eliminating those silent parts with amplitude below the threshold.

3.1.2 Audio Features

In this study, we extract 25 prosodic, 65 MFCC, and 15 formant frequency features. Prosody is mainly related to the rhythmic aspects of the speech, and believed to be the primary indicator of a speaker's emotional state [6]. The extracted prosodic features include: *Mean, Median, Standard Deviation, Max and Range of Pitch; Pitch Variation Rate; Rising/Falling Ratio; Max and Mean of Rising (Falling) Pitch Slope and Range; Overall Pitch Slope Mean, Median, and Standard Deviation; Mean, Median, Standard Deviation, Max and Range of Energy; Average Pause Length; and Speaking Rate.* Pitch is estimated based on cepstral analysis. Energy features are computed in time domain and represented in decibel (dB). Pitch variation rate is calculated as the division of total number of rise and fall over the number of segments. Speaking rate is the ratio of number and length of voiced segments.

MFCC and formant frequency are widely used in speech recognition, verification, and other applications. As our goal is to find out possible features that can truly represent the emotional state regardless of the speaker's cultural background, language, and accent etc., we also investigate these two types of features. The extracted MFCC features are the *mean, standard deviation, median, max, and min* of thirteen MFCCs of each utterance. Formant frequency features are the *mean, standard deviation, median, max, and min* of the first three formant frequencies. Formant frequencies are estimated by finding the locations of the resonances that make up an Infinite Impulse Response (IIR) filter obtained by using Linear Prediction Coding (LPC) [7]. In this paper, the spectral analysis is performed on speech segments of 512 points with 50% overlap.

3.2 Visual Feature Extraction

In this paper, the facial expression analysis is based on a key frame to represent the subject's emotional state in a video clip. The key frame is extracted as the frame at which the corresponding speech has the highest amplitude. We perform face detection first. A Gabor filter bank is then applied on this face image to extract visual features.

3.2.1 Face Detection

The face detection scheme that we applied is the planar envelope approximation method [8] in HSV color space. In this method, a pixel is considered as a skin pixel if the color of the pixel satisfies the following conditions:

$S \geq Th_s,\ V \geq Th_v\ ;\ S \leq -H-0.1V+110;\ H \leq -0.4V+75;$
If $H \geq 0,\ S \leq 0.08(100-V)\ H+0.5V$ ***Else*** $S \leq 0.5H+35,$

where Th_s and Th_v are set to 10 and 40, respectively.

After applying skin segmentation, some non-skin regions such as small isolated blobs and narrow belts are inevitably observed in the result as their color fall into skin color space. We apply morphological operation to implement the cleaning procedure. The detected face region is mapped back to the original image and normalized to a gray-level image of size 128×128 as the input to the Gabor filter bank.

3.2.2 Gabor Wavelet Features

Using Gabor wavelet features to represent facial expression has been explored in the literatures. It allows description of spatial frequency structure in the image while preserving information about spatial relations [3]. In this paper, we

employed Gabor filter dictionary design as described in [9]. The applied Gabor filter bank has 24 filters with 6 orientations and 4 scales. The mean and standard deviation of the coefficients of each of the filters are represented as the extracted features.

3.3 Feature Selection

We extracted 105 audio features and 48 visual features from each sample. However, not all the features can contribute to classification. Some of the features might even cause negative effects due to possible dependency among the features. Furthermore, with a feature space of 153 dimensions, the computational complexity is high. To reduce the dimensionality of the feature space, while maintaining the recognition accuracy, we performed feature selection using the stepwise method in SPSS (a trademark of SPSS Inc. USA). The criterion is Mahalanobis distance. The stepwise method starts from one feature. At each step, one feature is added to or removed from the selected feature subset to maximize the between-class Mahalanobis distance.

3.4 Classification

The classification scheme is based on Fisher's Linear Discriminant Analysis (FLDA). Linear Discriminant Analysis (LDA) assumes the discriminant function to be a linear function of data $\boldsymbol{x}$. In the case of c-class problem, the discriminant function is defined as:

$$g_i(x) = w_i^T x + w_{i0}, i = 1,2,...c, \quad (1)$$

In FLDA, the weights $\boldsymbol{w}$ are estimated by maximizing Fisher's criterion function $J_F(w)$. It can be deduced and found by solving the generalized eigenvalue problem:

$$J_F(w) = \frac{w^T S_b w}{w^T S_w w} = w^T S_b S_w^{-1} w, \quad (2)$$

where S_b and S_w are the between-class and within-class scatter matrix respectively. For classification, the input data is classified into the class that gives the greatest discriminant function value.

4. EXPERIMENTAL RESULTS

The experiments we performed were based on video samples from eight subjects, speaking six different languages. A total number of 500 samples, each delivered with one of six emotions were used for training and testing. From these samples, 360 samples (from six subjects) were selected for training, and the rest 140 (from the remaining two subjects) for testing. There was no overlap between the training and testing subjects.

4.1 FLDA *vs* Neural Network (NN)

In our first experiment, we performed classification on audio, visual, and audiovisual features separately. The applied FLDA classifier has six outputs corresponding to the six emotions (global classifier). A neural network (NN) architecture is also investigated for comparative study. The employed neural network is a three-layer feed-forward architecture based on back-propagation. As shown in Table 1, the combination of audiovisual features achieves better results than either of them only. FLDA performs better than neural network.

Classifier	Audio	Visual	Audiovisual
NN	44.29%	35%	45.71%
FLDA	66.43%	49.29%	70%

Table1: Recognition Results of FLDA and NN

4.2 Principal Component Analysis (PCA) *vs* Stepwise Method

In our second experiment, we performed feature selection by using PCA and stepwise method separately. PCA is a popular technique for dimensionality reduction. In PCA, the covariance matrix of the features is first computed. The eigenvalues and eigenvectors can then be calculated. These eigenvalues represent the contribution of each eigenvector to the total variation in the data. These eigenvectors are called the principal components. New data vectors can then be formed by projecting the original data onto the principal component vectors. In this study, we take the first m data from the newly generated data vectors, with m satisfy the following equation:

$$\sum_{i=1}^{m} E_i \Big/ \sum_{j=1}^{153} E_j \geq 90\%, \quad (3)$$

where E_i is the eigenvalue. This method achieves 62.86% accuracy, which is actually lower than before feature space reduction. As our goal is to reduce dimensionality, while maintaining or even achieving better accuracy, PCA is obviously not a good choice. We performed feature selection by using stepwise method. We selected 41 features from the original feature set, and the recognition rate was improved to 75.71%.

4.3 The Multi-Classifier Scheme

All the experiments that we described above are using global classifier. Feature selection and classification are all in a six-class output basis. However, different emotions could have different significant features, and the features to distinguish any combinations of the emotions also might be different. It will be desirable to find out the significant features for these scenarios.

We built six one-against-all (OAA) classifiers first, which is represented as "AN, DI, FE, HA, SA, SU" separately in Figure 1. Feature selection was performed to find significant features for individual class. For each OAA classifier, around ten to twenty features are selected. Compared with the 41 selected features by a global classifier, it is obvious that, for a specific emotion, some of the features selected in a global

scenario are redundant and might even cause negative effect. The output of each of these OAA classifiers is the probability of belonging to the corresponding emotion. We take this probability value as the input to a decision module for further classification.

We compared the performance of two rules in the decision module. In rule 1, if one of the outputs of these OAA classifiers is greater than 50%, we label the sample into the corresponding class. All the samples that have been misclassified, which means either none of the outputs beyond 50%, or two or more exceed 50%, will go to the global classifier for further classification. By using this method, the recognition result is improved to 79.29%.

In rule 2, we deal with the misclassified samples differently. If none of the outputs of OAA classifiers is greater than 50%, the sample will be further classified by a global classifier. If two or more of the outputs of the six OAA classifiers are greater than 50%, the sample will go to a separate classifier which is designed for those two or more specific emotional classes. Overall, we have built six OAA classifiers, 15 binary classifiers, 20 three-class classifiers, 15 four-class classifiers, six five-class classifiers, and one global classifier. In all these classifiers, feature selection is performed. This system achieves 82.14% accuracy, with the confusion matrix shown in Table 2.

		DETECTED					
		AN	DI	FE	HA	SA	SU
DESIRED	AN	**88.46**	0	3.85	3.85	0	3.85
	DI	0	**80.95**	4.76	9.52	4.76	0
	FE	0	18.18	**77.27**	4.55	0	0
	HA	0	16.00	0	**80.00**	4.00	0
	SA	0	4.76	14.29	0	**80.95**	0
	SU	4.00	0	8.00	4.00	0	**84.00**

Table 2: Confusion Matrix of the System (all in %)

4.4 Cross-validation Results

For the purpose of comparative study, we also performed experiments on a leave-one-out (LOO) cross-validation basis, which involves holding out one sample each time as the test data, and the rest of the samples as the training data. Stepwise method is also performed to select features. By using a global classifier, we achieve an overall accuracy of 89.2%. In this case, there is an overlap between the training and testing subjects, and thus the recognition rate is much higher. However, we can expect that, as more training subjects are added to the training set, the representation of human emotion will be better generalized toward a LOO cross-validation scenario, and better recognition accuracy can be achieved.

5. DISCUSSIONS AND CONCLUSIONS

In this paper, we proposed an emotion recognition system to recognize human affective state from audiovisual signals. The extracted features successfully captured the acoustic and visual characteristics of emotional data regardless of the user's cultural background, language, and race. Experimental results show that the combination of audio and visual information performs better than either of them only. We compared the performance of FLDA and NN. FLDA achieves much better results than NN, which reveals the linearity inherent of the proposed emotion recognition system. The applied feature selection algorithm efficiently reduces the dimensionality of the feature space, whilst achieving better recognition accuracy.

We proposed a classification scheme based on analysis of individual class and combinations of different emotions. Feature selection was performed to find significant features to distinguish emotions. This helps to solve the confusions between classes. However, as shown in Table 2, the major confusion has been between disgust and fear, sad and fear. From speech perspective, these emotions mostly have low amplitude and pitch variation. From visual perspective, the users tend to frown, and thus the Gabor features, which represent the texture information, can not distinguish the differences very well. In the future, we are going to investigate more distinctive features to distinguish different emotions. This analysis based on confusion matrix will surely help to achieve better accuracy.

6. REFERENCES

[1] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speech signals", *Eurospeech, Geneva, Switzerland,* Sep. 01-03, 2003

[2] M. W. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech", *Proceedings of 2004 IEEE International Symposium on Circuits and Systems,* Vancouver, May 23-26, 2004.

[3] M. J. Lyons, J. Budynek, A. Plante, and S. Akamatsu, "Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis", *Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition,* 28-30 March, 2000, Grenoble

[4] L. C. De Silva, T. Miyasato and R. Nakatsu, "Facial Emotion Recognition Using Multi-modal Information", *International Conference on Information, Communications and Signal Processing, ICICS '97,* Singapore, 1997.

[5] www.owlnet.rice.edu/~elec301/Projects01/dig_hear_aid/

[6] D. Hirst, "Prediction of prosody: An overview", *in Talking Machine: Theories, Models, and Designs. G. Bailly, C. Benoit, and T.R. Sawallis (editors). Elsevier Science Publishers, Amsterdam:* 1992

[7] www.phon.ucl.ac.uk/courses/spsci/matlab/

[8] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet parket analysis", *IEEE Transactions on Multimeida,* vol. 1, pp 264-277, Sept. 1999

[9] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, No. 8, August 1996