Recognizing Human Emotional State From Audiovisual Signals*

Yongjin Wang, *Student Member, IEEE*, and Ling Guan, *Fellow, IEEE*

***Abstract*—Machine recognition of human emotional state is an important component for efficient human-computer interaction. The majority of existing works address this problem by utilizing audio signals alone, or visual information only. In this paper, we explore a systematic approach for recognition of human emotional state from audiovisual signals. The audio characteristics of emotional speech are represented by the extracted prosodic, Mel-frequency Cepstral Coefficient (MFCC), and formant frequency features. A face detection scheme based on HSV color model is used to detect the face from the background. The visual information is represented by Gabor wavelet features. We perform feature selection by using a stepwise method based on Mahalanobis distance. The selected audiovisual features are used to classify the data into their corresponding emotions. Based on a comparative study of different classification algorithms and specific characteristics of individual emotion, a novel multiclassifier scheme is proposed to boost the recognition performance. The feasibility of the proposed system is tested over a database that incorporates human subjects from different languages and cultural backgrounds. Experimental results demonstrate the effectiveness of the proposed system. The multiclassifier scheme achieves the best overall recognition rate of 82.14%.**

***Index Terms*—Audiovisual information, emotion recognition, multiclassifier.**

I. Introduction

As COMPUTERS have become an integral part of our life, the need has arisen for a more natural communication interface between humans and machines. To accomplish this goal, a computer would have to be able to perceive its present situation and respond differently depending on that perception. To make human–computer interaction (HCI) more natural and friendly, it would be beneficial to give computers the ability to recognize situations the same way a human does.

In the field of HCI, audio and visual information are considered to be the two major indicators of human affective state, and thus play very important roles in emotion recognition. In this work, we explore methods by which a computer can recognize human emotion from audiovisual information. Such methods can contribute to human computer communication and to applications such as learning environment, entertainment, customer service, computer games, security/surveillance, and educational software [1].

Contemporary research in emotion in psychology and neurophysiology reveals that selected sets of so-called "basic" emotions constitute the foundations of human emotion. Certain emotions were associated with distinct facial signals, and these were common to cultures throughout the world [2]. A wide investigation on the dimensions of emotions reveals that at least six emotions are universal [3]. Several other emotions, and many combinations of emotions, have been studied but remain unconfirmed as universally distinguishable. A set of six principal emotions is: *happiness, sadness, anger, fear, surprise, and disgust*, which is the focus of study in this paper.

A great deal of studies has been conducted in machine recognition of human emotions in the past few years. The majority of these works either focus on speech alone, or facial expression only. However, as shown in [4], some of the emotions are audio dominant, while the others are visual dominant. When one modality fails or is not good enough to determine a certain emotion, the other modality can help to improve the recognition performance. The integration of audio and visual data will convey more information about the human emotional state. The complementary relationship of these two modalities will help to achieve higher recognition accuracy.

Recently, audiovisual based emotion recognition methods started to draw the attention of the research community. Song *et al.* [5], [6] extracted pitch and energy as audio features, and used the motion of eyebrow, eyelid, and cheek as expression features, while that of lips and jaw as the visual speech ones. The extracted three-stream audiovisual features were fused into a triple Hidden Markov Model (HMM) for classification. The proposed system was tested for a seven class problem (surprise, anger, joy, sad, disgust, fear, neutral), and around 85% recognition rate was claimed. De Silva and Ng [7] built an audio and a video system separately. In the audio system, pitch was extracted as the features and a nearest-neighbor method was used for classification. In the video system, they tracked the edge movement of lips, mouth corners, and eyebrows by using an optical flow algorithm, while HMM was trained as the classifier. A rule-based system was adopted to fuse the results of audio and video classification. Their system classified the six principal emotions and achieved a recognition rate 72% in a database that contains only two subjects. Chen *et al.* [8]

*Corrected. This paper first appeared in *IEEE Trans. Multimedia*, vol. 10, pp. 659–668. Due to a production error, the paper inadvertently appeared as a Correspondence.

Manuscript received September 26, 2006; revised January 13, 2008. Published July 9, 2008 (projected). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Horace Ho-Shing Ip.

Y. Wang is with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4 Canada (e-mail: ywang@comm.utoronto.ca).

L. Guan is with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3 Canada (e-mail: lguan@ee.ryerson.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2008.927665

utilized statistics of the pitch contour, energy envelope and their derivatives to represent the characteristics of emotional speech, and the visual information was obtained by tracking the position of eyebrow, cheek lifting, and mouth opening. By using the nearest mean classifier, their method produced 97.2% accuracy in classifying six principal emotions. However, the employed dataset only contains two speakers.

Chen *et al.* [9] performed facial expression analysis by tracking the movement of eyes, eyebrows, furrows and lips. Pitch contour, intensity contour, and energy spectrum were analyzed for acoustic feature representation. The extracted acoustic and facial features were combined and fed into the Support Vector Machine (SVM) for emotion recognition. They classify the six principal emotions with neutral state on a two-speaker dataset, and achieved an average of 84% accuracy. Hoch *et al.* [10] introduced a person dependent emotion recognition system to be applied in an automotive environment. The acoustic signals were recognized by a neural network approach on statistics of low level features such as pitch, power, formants and duration of voiced segments. Facial expressions were evaluated by a SVM classifier on Gabor-filtered face regions. The bimodal fusion was performed at the decision level through a weighted linear combination. Their system classifies three emotions (positive, negative, neutral) with an average of 90.7% recognition rate obtained. Zeng *et al.* [11] extracted log energy, syllable rate, and pitch as prosodic features, and tracked a set of predefined action unit for facial expression representation. A smoothing method was applied to reduce the influence of speech on facial expressions. Bimodal fusion was performed at the decision level using a voting method. In another paper [12], similar feature extraction was performed as in [11], while a Hidden Markov Modal (HMM) based approach was used to fuse audio and visual streams. Their proposed approach produced 85.24% recognition rate in classifying six principal emotions and neutral state.

In summary, existing bimodal approaches demonstrate that the performance of emotion recognition systems can be improved by integrating audio and visual information. However, it is far from a solved problem due to limits in the accuracy and generality of proposed systems. In particular, most of the existing solutions only focus on recognizing human subjects that speak a single language. In this paper, we propose a system to recognize human emotional state from audiovisual information. The proposed system is tested over a database which features language, speaker and context independence. Audio and visual features are first extracted to represent the vocal and facial characteristics of humans in different emotions. We perform feature selection to find out significant features, whilst reducing the dimensionality of the feature space. To achieve higher recognition rate, we compare different classification algorithms and select one which is most suitable for the recognition problem on hand. Furthermore, as different emotions might have different significant features, and the features to distinguish combinations of different emotions might also be different, we propose a novel multiclassifier scheme to analyze these scenarios.

The remainder of this paper is organized as follows: Section II presents the data collection method. The audio and visual feature extraction methods are discussed in Section III and IV respectively. Section V presents the emotion recognition system and the proposed multiclassifier scheme, along with the detailed experimental results. Finally, conclusions are drawn in Section VI.

II. Data Acquisition

The performance of an emotion recognition system is highly dependent on the quality of emotional data on which it is trained. When working with speech and facial expression, special care must be taken to ensure that the particular emotion is properly vocalized and expressed. We set up an audiovisual data collection system to record emotional video data. A digital video camera was used to record the samples in a relatively quiet and bright environment. Our experimental subjects were provided with a list of emotional sentences and were directed to express their emotions as naturally as possible by recalling the emotional incident, which they had experienced in their lives. To ensure the context independency of the speech data, we provided more than ten reference sentences for each emotional class. The list of emotional sentences was provided for reference only. Every language has a different set of rules that govern the construction and interpretation of words, phases, and sentences. While some subjects expressed their emotions by using the same sentence structure as provided, others opted to use variations or different sentences according to their cultural background.

For the purpose of a more general application, the data should not be restricted to the user's language, accent, and cultural background. To ensure the diversity of the database, we collected video samples from eight subjects, speaking six different languages. The six languages are English, Mandarin, Urdu, Punjabi, Persian, and Italian. Different accents of English and Mandarin were also included. Some of the subjects have facial hair, which further increases the diversity of the database. To ensure the correct expression of human emotion, the experimental dataset was selected based on listening test by at least two participating human subjects who do not know the corresponding language. A video sample was added to the experimental dataset if and only if all testing subjects perceive the intended emotion. For English language, a sample was selected based on the correct perception of all the eight subjects. We collected a total of 500 video samples, each delivered with one of the six particular emotions. The clips were recorded at a sampling rate of 22050 Hz, using a single channel 16-bit digitization.

III. Audio Feature Extraction

To build an emotion recognition system, the extraction of features that can truly represent the universal characteristics of the intended emotion is a challenge. For emotional speech, a good reference model is the human hearing system. Previous works have explored several different types of features. As prosody is believed to be the primary indicator of a speaker's emotional state [13], most of the works adopt prosodic features [14]–[16]. However, Mel-frequency Cepstral Coefficient (MFCC) and formant frequency are also widely used in speech recognition and some other speech processing applications, and have also been studied for the purpose of emotion recognition [17]–[19]. As our goal is to simulate human perception of emotion, and identify possible features that can convey the underlying emotions in speech regardless of the language, speaker, and context, we investigate all these three types of features.

A. Preprocessing

The collected emotional data usually contain noise due to the background and "hiss" of the recording machine. Generally, the presence of noise will corrupt the signal, and make the feature extraction and classification less accurate. In this work, we perform noise reduction by thresholding the wavelet coefficients [20]. Leading and trailing edges are then eliminated since they do not provide useful information. To perform spectral analysis for feature extraction, the preprocessed speech signal is segmented into speech frames using a Hamming window of 512 points with 50% overlap.

B. Prosodic Features

Prosody is mainly related to the rhythmic aspects of the speech, and is normally represented by the statistics and variations of fundamental frequency, intensity, speaking rate, etc. In this work, we extracted 25 prosodic features as listed in Table I.

TABLE I
LIST OF EXTRACTED PROSODIC FEATURES

Index	Feature Description
1	Pitch Mean,
2	Pitch Median
3	Pitch Standard Deviation
4	Pitch Max
5	Pitch Range
6	Pitch Variation Rate
7	Rising/Falling Ratio
8	Rising Pitch Slope Max
9	Falling Pitch Slope Max
10	Rising Pitch Slope Mean
11	Falling Pitch Slope Mean
12	Pitch Rising Range Max
13	Pitch Falling Range Max
14	Pitch Rising Range Mean
15	Pitch Falling Range Mean
16	Overall Pitch Slope Mean
17	Overall Pitch Slope Standard Deviation
18	Overall Pitch Slope Median
19	Energy Mean (dB)
20	Energy Median (dB)
21	Energy Standard Deviation (dB)
22	Energy Max (dB)
23	Energy Range (dB)
24	Average Pause Length
25	Speaking Rate

The pitch is estimated based on the Fourier analysis of the logarithmic amplitude spectrum of the signal [21]. The energy features are extracted in time domain and represented in decibel (dB). Pitch variation rate R_{var} and pitch rising/falling ratio R_{rf} are calculated respectively as

$$R_{\text{var}} = \frac{N_{rise} + N_{fall}}{N_{frame}}, \quad R_{rf} = \frac{N_{rise}}{N_{fall}} \tag{1}$$

where N_{frame} is the number of speech frames, N_{rise} and N_{fall} are the number of speech frames with continuous rising and falling pitch respectively.

Speaking rate is approximated by

$$R_{spk} = \frac{1}{mean\_segment\_length} = \frac{N}{\sum_{i}^{N} T_i} \tag{2}$$

where T_i is the length of voiced segment i and N is the number of voiced segments. The voiced segments are defined as the segments of speech signal between pauses.

Pitch slope of each rise and each fall is calculated as

$$S_{pitch} = \frac{pitch\_difference}{rise(fall)\_length} = \frac{|f_{\max} - f_{\min}|}{t_{end} - t_{start}} \tag{3}$$

where f_{max} and f_{min} denote the maximum and minimum pitch value on the rise (fall) respectively. t_{start} and t_{end} represent the starting and ending time of the rise (fall).

Pauses are only used to compute Average Pause Length, and they are discarded when computing other parameters, which are focused on voiced signal. Pitch (amplitude) range is determined by scanning the curve, finding the maximum and minimum pitch (amplitude), and calculating the difference.

C. MFCC Features

Mel-frequency Cepstral Coefficient (MFCC) is a popular and powerful analytical tool in the field of speech recognition. The purpose of MFCC is to mimic the behavior of human ears by applying cepstral analysis. In this paper, the implementation of MFCC feature extraction follow the same procedure as described in [22]. The MFCCs are computed based on speech frames. However, the lengths of the utterances are different, and thus the total number of coefficients is different. In order to facilitate classification, the features of each utterance mapped to the feature space should have the same length. Furthermore, with a feature vector of high dimension, the computational cost is high. Usually, in speech recognition, the total number of coefficients being used is between nine and thirteen. This is because most of the signal energy is compacted in the first few coefficients due to the properties of the cosine transform. In this work, we take the first 13 coefficients as the useful features. We then calculate the mean, median, standard deviation, max, and min of each order of coefficients as the extracted features, which produce a total number of 65 MFCC features.

D. Formant Frequency Features

Formant frequencies are the properties of the vocal tract system. In this paper, the formant frequency estimation is based on modeling the speech signal as if it were generated by a particular kind of source and filter [21]. To find the best matching system, we use the Linear Prediction method. In order to make the size of the formant frequency features uniform, and achieve compromise between the imitation efficiency of the vocal tract system and dimensionality of the feature space, we take the mean, median, standard deviation, max and min of the first three formant frequencies as the extracted features. In this way, we extract a total number of 15 formant frequency features from each utterance.

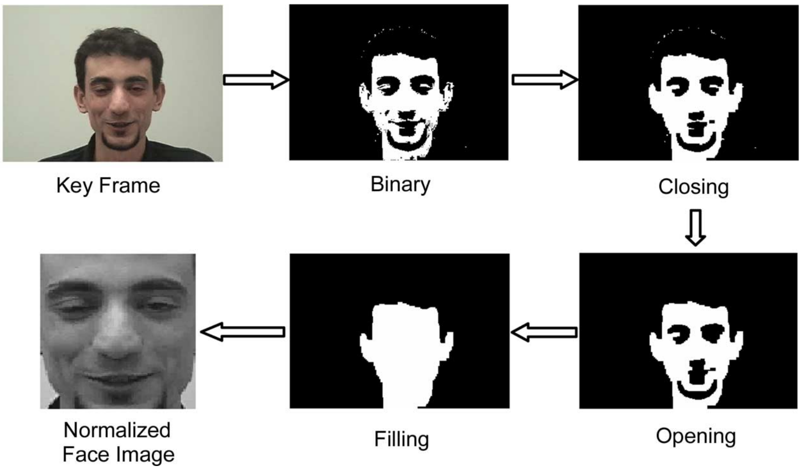


Fig. 1. Procedure of the applied face detection scheme.

IV. Visual Feature Extraction

Facial expression is another major factor in human emotion recognition. In general, the face region is detected from the image first, and then facial expression information can be extracted from the observed facial images or image sequences. In the case of still image, extracting facial expression information means to localize the face and its features from a single image [23]–[26]. In the case of image sequence, it means to track the motion of the face and its features in the image sequence [27], [28]. Although the latter case may provide more accurate facial representation, it generally requires more computation. In this paper, we use a key frame to represent the subject's emotional state in a video clip, where the key frame is extracted as the one with the highest speech amplitude. The underlying idea for selecting key frame is based on intuition and observation that human facial features will be exaggerated at large voice amplitude.

Existing solutions for facial expression analysis can be roughly categorized into two groups. One is to treat the human face as a whole unit [23], [24], and the other is to represent the face by prominent components, such as the mouth, eyes, nose, eyebrow, and chin [25]–[28]. The analysis of facial components is critically dependent on the accurate localization of the local features. Further, focusing on only a few facial components, the representation of the discriminant characteristics of human emotion might be inadequate. In this work, we perform facial analysis by treating the face as a holistic pattern. A face detect scheme based on HSV color model is then used to detect the face from the background. The visual information is represented by Gabor wavelet features.

A. Face Detection

Different approaches of face detection have been studied in the past. Examples of these approaches include Principal Component Analysis [25], skin color analysis [26], and filtering techniques [27]. Among these methods, color analysis has been quite popular due to its effectiveness and fast processing speed. The face detection scheme that we used in this study is the Planar envelope approximation method [29] in HSV color space. 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 operations to implement the cleaning procedure. As shown in Fig. 1, the detected face region is mapped back to the original image, and cropped such that the major components of the face are included. The cropped face region is normalized to a gray-level image of size 128 × 128 as the input to the Gabor filter bank.

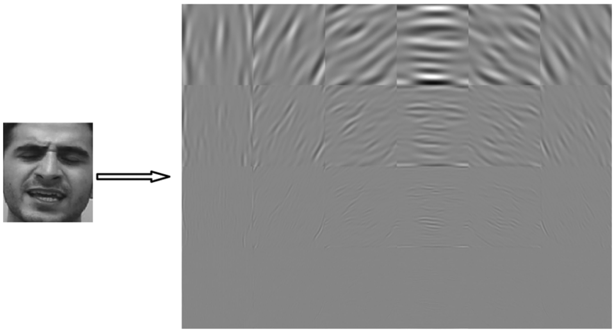


Fig. 2. Example of Gabor wavelet transformed image.

B. Gabor Wavelet Representation

Using Gabor wavelet features to represent facial expressions have been explored and shown to be very effective in the literature [23]. It allows description of spatial frequency structure in the image while preserving information about spatial relations. In this paper, the Gabor filter bank is designed using the algorithm proposed in [30]. The designed Gabor filter bank consists of filters in 4 scales and 6 orientations. Fig. 2 shows an example of Gabor wavelet transformed face image. For an input image of size of 128 × 128, a total number of $128 \times 128 \times 4 \times 6 = 393216$ Gabor coefficients are generated. With a feature space of such high dimensionality, the computational cost is very high, and thus it is not very suitable for practical applications. We therefore consider the mean and standard deviation of the magnitude of the transform coefficients of each filter as the features. This results in a feature vector of 48 dimensions.

V. Emotion Recognition System

We extract 105 audio and 48 visual features from each emotional audiovisual sample. Fig. 3 sketches an overview of the proposed recognition system. An input video sequence is passed

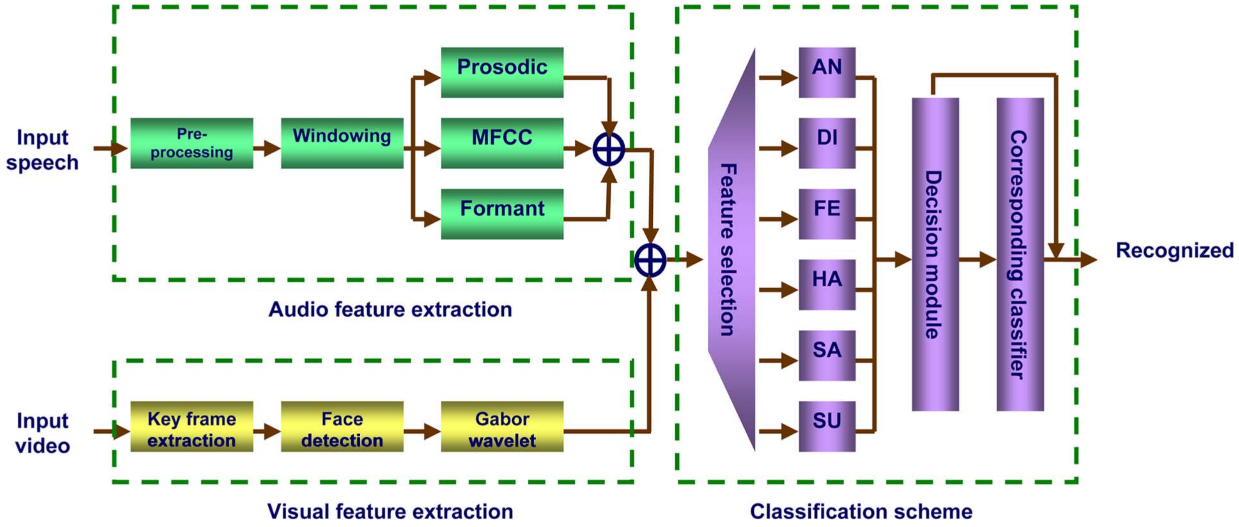


Fig. 3. Overview of the emotion recognition system.

through two different channels to extract the audio and visual features. These two streams of features are then concatenated into one stream, and a feature selection process is applied to find out significant features. The system is trained by using the known-class data, and a new input can be classified into the corresponding class by using the adopted classification scheme.

The experiments that we performed were based on video samples from eight subjects, speaking six languages. A total number of 500 samples were used for training and testing, each delivered with one of the six principal emotions. The average duration of each sample is around 5 seconds. Since the samples were recorded in audiovisual format, the input to the audio and visual feature extraction channel has the same length in terms of time duration. During classification, the correspondence between audio and visual features is taken into consideration by proper labelling in accordance with the original video samples. From these samples, 360 were selected for training, and the other 140 for testing. There is no overlap between the training and testing subjects. Each emotion has approximately the same number of samples for training and testing. In this section, we first present the experimental results of applying different classification algorithms. We then compare feature selection algorithms by using the best-performance classifier. A multiclassifier scheme based on the analysis of individual class and combinations of different classes is then introduced.

A. *Comparison of Classification Algorithms*

To classify the extracted features into different human emotions, we need to select a classifier that can properly model the data and achieve better classification accuracy. Since we do not have any prior knowledge about the characteristics of the features, a comparison of popular classification algorithms in emotion recognition will help us achieve insights into the problem and select an appropriate method to build upon. In this work, we compare the performance of Gaussian Mixture Model (GMM), k-nearest Neighbors (k-NN), Neural Network (NN), and Fisher's Linear Discriminant Analysis (FLDA).

GMM models the probability density function (pdf) of the data as weighted sum of several different Gaussian density functions. Expectation Maximization (EM) algorithm can then be used to estimate the parameters of GMM, including probability, mean, and covariance matrix of each component. For classification, GMM is usually performed in a modular architecture, which involves a separate GMM being trained for each individual class. An input signal is labeled with the class corresponding to the maximum output. Since the exact distribution of the data is not known a priori, in this paper, the number of Gaussian functions $k = 8$ is selected based on empirical analysis.

k-NN is a nonparametric method for classification [31]. It assigns a class label to the unknown data by examining its k nearest neighbors of known data $\mathbf{x}$. For a new input vector $\mathbf{y}$, the k-NN algorithm identifies a subset of k feature vectors from the reference training samples that lie closest to $\mathbf{y}$ with respect to a pattern similarity function $\mathrm{D}(\mathbf{y}, \mathbf{x})$, usually Euclidean distance. A majority voting is then applied to determine the corresponding class. A popular way to determine the k value is to use leave-one-out cross validation. However, the k value selected using this method might not be the best k value. In this work, we perform experiments on ten k values from one to ten, and the best result is achieved when $k = 7$.

A three layer feedforward neural network [32] is also investigated for classification. Unlike the classical statistical methods such as the Bayes classifier, no knowledge of the underlying probability distribution is needed by a neural network. It can learn the free parameters, weights and biases, through training of samples. The number of input layer neurons is equal to the dimension of the input feature set, while the output neurons corresponding to the six emotion classes. The number of hidden layer neurons is determined by $N_H = \sqrt{N_I N_O}$, where N_I and N_O are the number of input and output neurons respectively. Backpropagation algorithm is used to train the network. The number of epochs for neural network training is 150. A new input is labeled the class that produces maximum output value.

TABLE II
COMPARISON OF CLASSIFICATION ALGORITHMS

Classifier	Prosodic	Audio	Visual	Audiovisual
GMM	50.97%	56.97%	21.68%	65.38%
K-NN	49.29%	61.43%	29.29%	62.86%
NN	49.29%	51.43%	35.00%	56.43%
FLDA	65.71%	66.43%	49.29%	70.00%

TABLE III
COMPARISON OF FEATURE SELECTION ALGORITHMS

	Dimension	Recognition Results
Original feature set	153	70.00%
PCA	40	62.86%
Stepwise method	45	75.71%

Linear discriminant analysis [31] assumes the discriminant function $\mathrm{g}(\mathbf{x})$ to be a linear function of data $\mathbf{x}$. In the case of c-class problem, the discriminant function is defined as $g_i(x) = w_i^T x + w_{io}$, where w_i is a vector, and w_{i0} is a constant. Fisher's LDA finds a set of M basis vectors w by maximizing the ratio of between-class and within-class scatter matrices. For classification, the input data is classified into the class that gives the greatest discriminant function value.

Table II shows the experimental results of applying different classifiers on 25 prosodic features, 105 audio features, 48 visual features, and 153 audiovisual features separately.

The experimental results show that the combination of audio and visual information performs better than either of them only. FLDA outperforms the other classifiers. GMM and k-NN are statistical methods that are based on the estimation of probability density function. Neural network estimate the weights and bias values between different layers through a training process. They all need the training set to be large enough for more accurate estimation. In our case, the training set contains 360 samples, and the dimension of the feature space is 153. The high dimensionality of the feature space and the sparseness of the training set limit the accuracy of the estimation, and thus the performance.

In our experiments, prosodic features are confirmed to be a powerful indicator of human emotional state in speech. By using FLDA, 25 prosodic features produce a recognition accuracy of 65.71%, which is very close to 66.43%, the results of using all the audio features. However, this also demonstrates that phonetic features also contribute to classification.

Another observation is that the complementary relationship of audio and visual information enhances the performance of the system. Fig. 4 provides a graphical comparison of the confusion matrix of applying FLDA on audio, visual, and audiovisual features. It is obvious that *anger, disgust, sadness* and *surprise* can be better distinguished in audio, while *fear* and *happiness* are more accurately recognized by the visual cues. By combining audio and visual features, the recognition accuracy of *disgust, fear, happiness*, and *sadness* are improved. However, the classification accuracy of *anger* and *surprise* are not as good as before the integration of data. As shown in Fig. 4, *anger* and *surprise* were not well recognized by using visual information. This demonstrates that the representation of visual information is insufficient and caused negative effects to the integration. A feature selection process is therefore needed to deal with this problem.

B. Feature Selection: PCA versus Stepwise Method

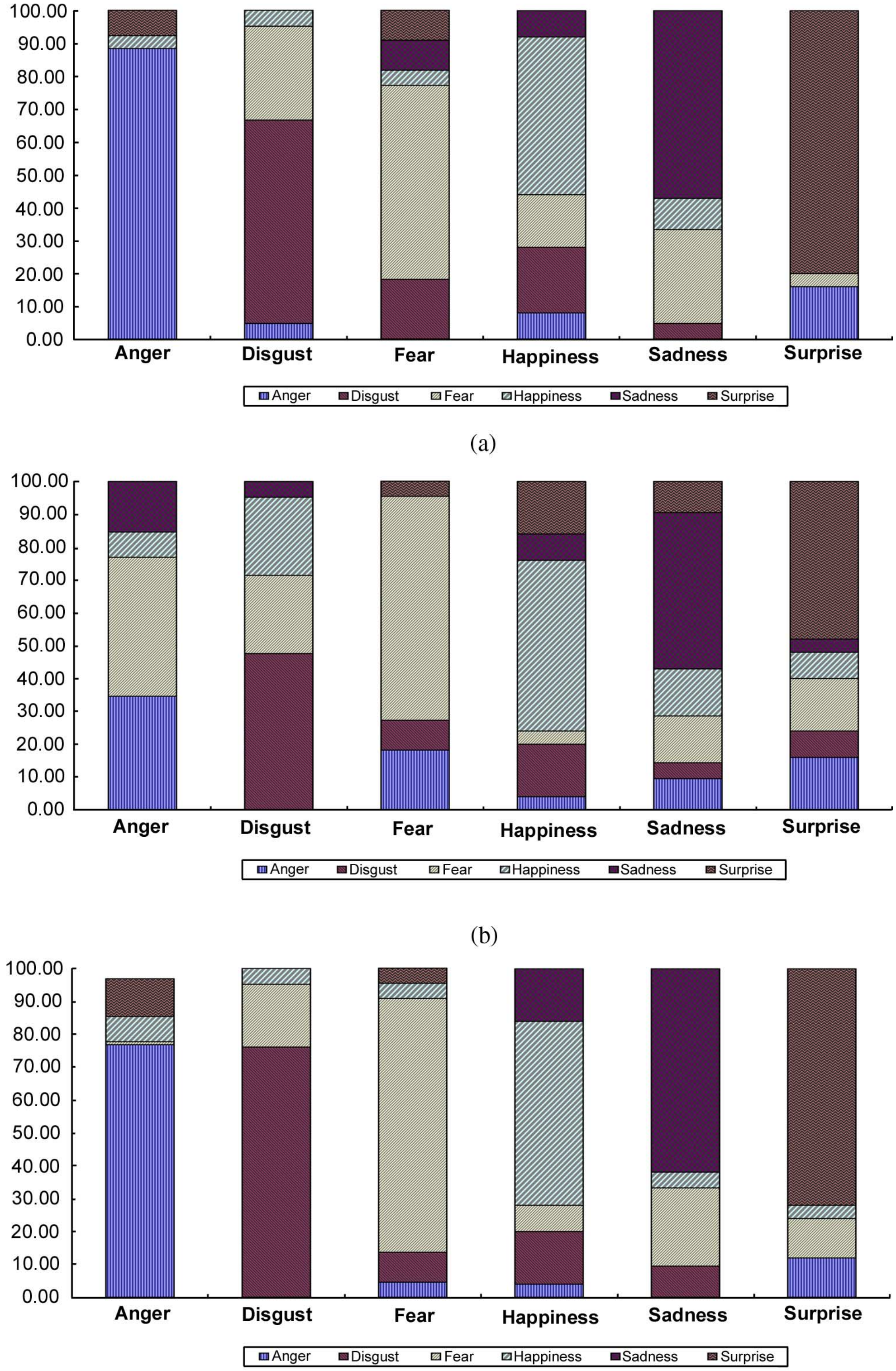
The performance of a pattern recognition system critically depends on the discriminant ability of the features in terms of separating patterns belonging to different classes in the feature space. The importance of selecting relevant subset from the original feature set is closely related to the "curse of dimensionality" problem in function approximation. In this problem, sample data points become increasingly sparse as the dimensionality of the function domain increases, such that the finite set of samples may not be adequate for characterizing the original mapping. In addition, the computational requirement is greater for implementing a high-dimensional mapping. To alleviate these problems, we need to reduce the dimensionality of the input domain.

Many different feature selection and dimensionality reduction methods exist, which can be roughly divided into two categories. In one of them, a subset of the original features is selected. The selected subset of features retains the original feature characteristic. In another category, the feature space is transformed into another domain. The discriminant information is concentrated in part of the coefficients in the transformed domain. Dimensionality reduction can be achieved by truncating the newly generated transform-domain features. In this paper, we study PCA and the stepwise method as representatives of each category.

PCA reduces the dimensionality of the data by performing eigen-analysis on the covariance matrix of the original data [32]. The covariance matrix is a square matrix, and thus the eigenvectors and associated eigenvalues can be calculated. Ordering the eigenvectors by sorting the associated eigenvalues from the highest to the lowest gives the components in order of significance. The components with less significance can be ignored and thus dimension reduction can be achieved. In PCA parameter determination, we use a criterion by taking the first M eigenvectors that satisfy $\sum_{i=1}^{M} \lambda_i / \sum_{i=1}^{N} \lambda_i \geq 90\%$, where λ_i is the eigenvalue, N is the dimension of feature space.

The stepwise method is implemented in SPSS (a trademark of SPSS, Inc., USA). It starts with one feature and progressively adds one feature at a time. The measure to determine the inclusion and exclusion of a feature is the Mahalanobis distance. For each step, one feature is added to or removed from the selected feature subset to maximize the between-class Mahalanobis distance. In our experiment, we selected 45 features from the original feature set, with the corresponding index numbers listed in Table IV.

Table III reports the dimensionality and recognition accuracy after applying the two feature selection algorithms. The recognition accuracy after applying PCA is actually lower than before feature space reduction. This shows the loss of information during the PCA transformation. As our goal is to reduce dimensionality, whilst maintaining or even achieving better accuracy, PCA is obviously not a good choice. The stepwise method reduces the dimensionality, and the recognition rate is also improved significantly.



(a)

(b)

(c)

Fig. 4. Graphical demonstration of the confusion matrix by applying FLDA classifier on (a) audio, (b) visual, and (c) audiovisual features.

In comparison with the original feature set, the selected feature subset achieves better overall accuracy, specifically in *anger, happiness, sadness*, and *surprise* (Fig. 5). However, it can be observed that the performance on *disgust* is actually worse than without selection. This is because the stepwise method is a suboptimal feature selection algorithm, and thus an optimal feature subset can not be guaranteed. Furthermore, the feature selection method is implemented in a global scenario, by which the selected subset is to distinguish all the six classes. However, different emotions could have different significant features that separate them from all the other emotions or some specific emotions. This inspires us to consider a new proposal: performing feature selection and classification on individual emotion and combination of different emotions.

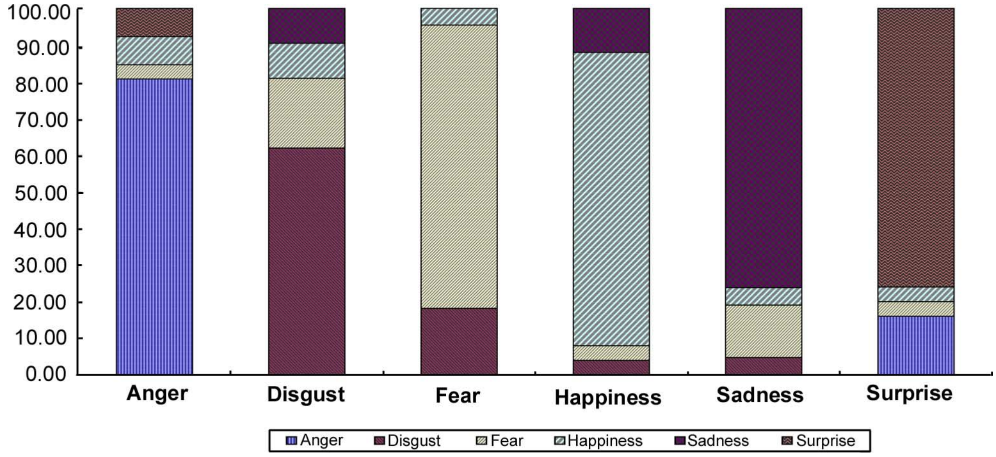


Fig. 5. Graphical demonstration of the confusion matrix by applying FLDA classifier on stepwise method selected audiovisual features.

C. Multiclassifier Scheme

The key goal of a multiclassifier system is to obtain a better composite global model, with more accurate and reliable estimates or decisions [33]. One approach in the design of a multiclassifier system is to combine the outputs of individual classifiers, where each classifier solves the same classification problem. Each classifier may use different subsets of the training data, and may use different feature extractors. The outputs of individual classifiers are combined through certain rules such as voting, averaging, and product rule. A number of research works that analyze such ensemble based methods can be found in [34]–[36]. In general, ensembles can improve the recognition results, but it requires a large number of training samples and is computationally complex.

A central consideration in the design of our classification scheme is to decompose a complex global emotion recognition task into a set of simpler local emotion recognition subtasks based on the so called "divide and conquer" principle [33]. In pattern recognition, when the number of classes is large, the boundaries between different classes tend to be complex and hard to separate. It will be easier if we can reduce the possible number of classes and perform classification in a smaller scope. In this paper, a multiclassifier scheme involving the analysis of individual class and combinations of different classes is proposed. As FLDA has shown its effectiveness in the previous experiments, the individual classifiers in this multiclassifier scheme are based on FLDA. The architecture of the multiclassifier scheme is shown in Fig. 6.

We built six one-against-all (OAA) classifiers first, which are represented as "AN, DI, FE, HA, SA, SU" separately in Fig. 6. The OAA classifiers are designed specifically for individual emotion, with each of them performs a 2-class pattern recognition problem. In the training process, for each OAA classifier, we label all the samples that do not belong to the corresponding emotion as one class. The output of each of these OAA classifiers is the probability of belonging to the corresponding emotion. For example, in the "AN" OAA classifier, all the samples of *anger* are labeled as "*anger*," while all the other samples are labeled as "*non-anger*." If the output value is greater than 50%, the sample may be classified as "*anger*," and otherwise "*non-anger*." In each OAA classifier, feature selection was performed to find significant features for the corresponding emotion. The applied feature selection algorithm is the stepwise method, which has been shown to perform better in Section V-B. The selected features for individual emotion (OAA classifier) are listed in Table V.

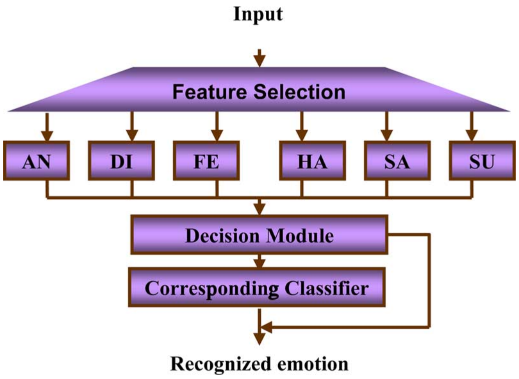


Fig. 6. Multiclassifier scheme.

The output of each OAA classifier is taken as the input to a decision module for further classification. We compared the performance of two rules in the decision module. Let N be the number of OAA classifiers whose output exceeds 50%, we have

Rule 1:

If $N = 1$,

Then label as the corresponding emotion

Else ($N = 0$ *or* $N >= 2$) go to a global six-class classifier

Rule 2:

If $N = 1$,

Then label as the corresponding emotion

Else if $\mathrm{N} = 0$

Then go to a global six-class classifier

Else $(N >= 2)$ go to specific N-class classifier

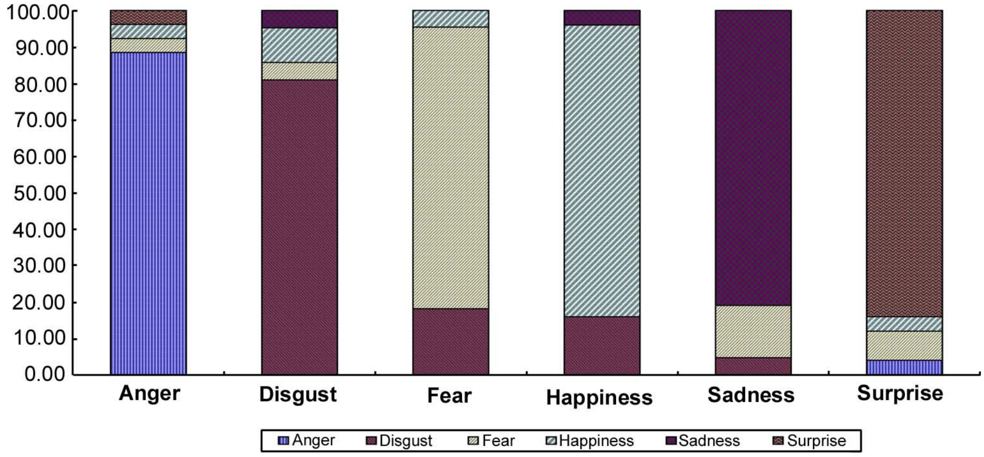


Fig. 7. Graphical demonstration of the confusion matrix by applying the proposed multiclassifier scheme.

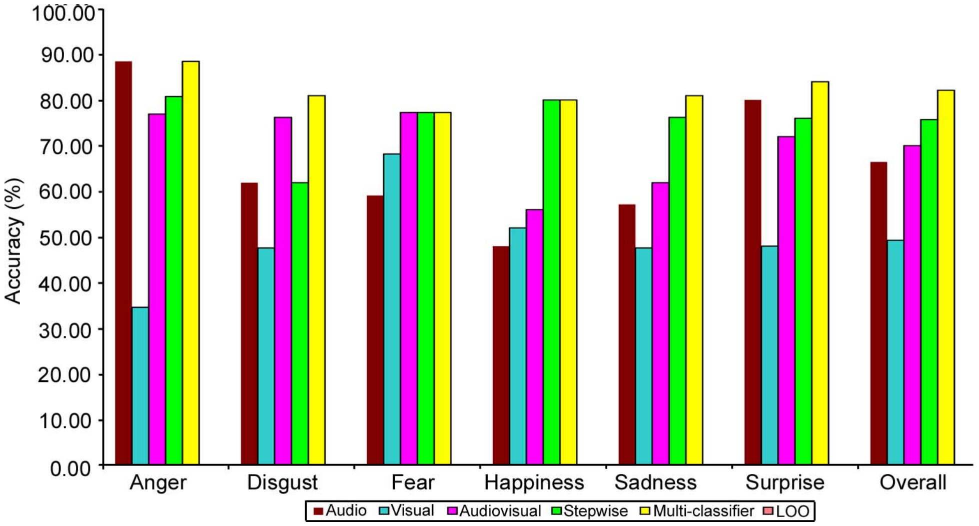


Fig. 8. Comparison of recognition results.

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 exceeds 50%, or two or more are greater than 50%, will go to the global classifier for further classification. By using this rule, the recognition results are 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%, which means the sample might belongs to more than one emotion, the sample will go to a separate classifier which is designed for those two or more specific emotional classes. The underlying reason for us to select this scheme is that, there might be specific set of features that can better distinguish two or more different emotions. If a sample possesses strong characteristics of two or more emotions, a classifier dedicated to distinguishing minor differences in these emotions will help to make a correct decision. Overall, we have built six OAA classifiers, 15 binary classifier, 20 three-class classifiers, 15 four-class classifiers, six five-class classifiers, and one global classifier. In total, there are 63 classifiers, and the stepwise method is used to select the appropriate input features for each of them. This system achieves 82.14% accuracy, with the graphical demonstration of the confusion matrix shown in Fig. 7.

TABLE IV
SELECTED GLOBAL FEATURES USING STEPWISE METHOD

	Feature Index Number
FLDA	[2 3 7 19 21 22 24 25 26 27 28 34 43 48 55 68 78 80 91 107 108 109 111 113 114 116 118 119 121 122] 124 127 128 130 131 132 133 134 135 137 139 143

From Table V we can find that the selected significant features for individual emotions are different. Some of the features selected in a global scenario are redundant (Table IV), and some of the other features might contribute to the classification of specific emotion. In all the cases, the selected subsets incorporate both audio and visual features. This explains that each emotion is associated with certain audio and visual characteristics, and has a specific set of attributes to distinguish it from others. The recognition performance can be improved by successful localization of these attributes.

TABLE V
SELECTED FEATURES FOR INDIVIDUAL OAA CLASSIFIERS

OAA	Feature Index Number
AN	[1 7 13 19 24 34 35 49 66 71 72 80 89 106 108 119 121 124 125 128 132 138 143 145 148 151 152 153]
DI	[4 20 24 25 33 43 54 60 61 74 88 106 107 108 113 116 122 134 135 137 149]
FE	[1 2 14 24 38 39 55 65 67 108 113 118 127 132 133]
HA	[1 3 6 32 49 50 55 56 65 74 75 78 80 98 99 108 109 111 116 122 124 128 130 131 132 133 135 138 142 145]
SA	[1 2 14 22 24 25 29 35 37 58 86 103 106 111 113 118 135 143 146]
SU	[3 7 17 18 22 25 34 52 58 79 89 107 109 113 115 119 122 130 132 134 135 141 144 145 146 147 149 151]

(Notes: Indices 1–25 correspond to prosodic features, 26–90 correspond to MFCCs, 91–105 are formant frequency features, and 106–153 are visual Gabor features)

Another interesting observation is that there is not even a single feature which is significant for all the classes. This actually reveals that nature of human emotion is such that there are no sharp boundaries between emotions. One emotion might have similar patterns with some of the other emotions, whilst having different patterns with the rest. The human perception on emotion is based on the integration of different patterns. For example, sadness and disgust both have long mean pause length (index: 24) and low tempo (index:25) to distinguish then from the other four emotions, and these two features are selected. Then, sadness can be further separated from disgust because it has lower energy max (index:22).

By performing individual class based analysis, the recognition rate improves significantly. In our proposed multiclassifier scheme, Rule 2 achieves higher accuracy than Rule 1. This demonstrates that the significant features for distinguishing different combinations of emotions are different. By using Rule 2, our system achieves the best overall accuracy, and best recognition rate for all the individual emotions.

It should be noted that although the proposed multiclassifier scheme involves a large number of classifiers, but the training process of these classifiers are assumed to be offline. The computational complexity at the recognition stage involves at most two steps, the computation of binary classifiers, and the corresponding n-class classifier (which is dependent on the output of first level classification), therefore do not increase the computational cost significantly.

D. Cross-Validation

For the purpose of comparative study, we also performed experiments on a leave-one-out (LOO) cross-validation basis. LOO cross-validation works as follows: for each time, one sample is held out as the testing data, while the rest of the data in the entire data set is used as the training data. This procedure continues until all the individual samples in the entire data set have been held out once. The recognition accuracy is calculated as the ratio of the number of correctly classified samples and the total number of samples in the data set. We perform feature selection using the stepwise method. By using a global classifier, we achieve an overall accuracy of 89.2%.

VI. CONCLUSIONS

In this paper, an audiovisual based emotion recognition system is presented. Extensive experiments were conducted to test the effectiveness of our system. The recognition results using different features and a classification scheme based on Fisher's Linear Discriminant Analysis (FLDA) are summarized in Fig. 8. The results demonstrate that the combination of audio and video information performs better than either one alone. The applied stepwise method efficiently reduces the dimensionality of the feature space, whilst achieving better recognition accuracy. The proposed multiclassifier scheme produces noticeable improvement in individual class recognition accuracy, and achieves the best overall recognition rate of 82.14%. This multiclassifier scheme takes advantage of the analysis of significant features in an individual class and uses such to distinguish any combinations of any classes. It helps to obtain more detailed insight into individual emotion and the way to separate specific emotions. This "divide and conquer" method partitions classification into finer analysis, and is a popular practice in pattern recognition problems.

By using leave-one-out cross-validation method, we achieve very promising results of 89.2% correct recognition rate. This demonstrates that the extracted features successfully captured the vocal and visual characteristics of emotional data regardless of the user's cultural background and language. In the case of LOO, there is overlap between the training subjects and testing subjects, and thus the recognition rate is 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 expected.

Considering a more generic application, our system was tested using a versatile database, in which samples from different subjects, speaking different languages were collected. By performing experiments on such a dataset, we conducted an exploring study of recognizing human emotions independent of language, speaker, and context. Language and cultural background might have some influence on the way in which people express their emotions, our proposed system retains good results, and demonstrated the possibility that the emotional expressions can be identified beyond these boundaries. Although by no means conclusive, the presented study sheds some interesting lights to this important aspect of human emotion recognition.

In our experiments, visual feature based classification accuracy is low. This demonstrates that the visual feature representation is not strong enough. In the future, we aim to work on a hybrid approach which incorporates the analysis of the whole face, the prominent facial components (such as mouth, eyes, etc.), as well as dynamic features in video sequence. Data fusion algorithms will be further investigated to better utilize the complementary relationship of the two modalities, so as to improve the efficiency of the system.

REFERENCES

[1] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[2] P. Ekman and M. O'Sullivan, "The role of context in interpreting facial expression: Comment on Russell and Fehr," *J. Exper. Psychol., General*, vol. 117, pp. 86–88, 1987.

[3] P. Ekman, "Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique," *Psychol. Bull.*, vol. 115, pp. 268–287, 1994.

[4] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proc. IEEE Int. Conf. Information, Communications and Signal Processing*, Singapore, Sep. 1997, vol. 1, pp. 397–401.

[5] M. Song, C. Chen, and M. You, "Audio-visual based emotion recognition using triples hidden Markov model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada, May 2004, vol. 5, pp. 877–880.

[6] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition: A new approach," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 1020–1025.

[7] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, France, Mar. 2000, pp. 332–335.

[8] L. S. Chen, H. Tao, T. S. Huang, T. Miyasato, and R. Nakatsu, "Emotion recognition from audiovisual information," in *Proc. IEEE 2nd Workshop on Multimedia Signal Processing*, CITY?, CA, Dec. 1998, pp. 83–88.

[9] C. Chen, Y. Huang, and P. Cook, "Visual/acoustic emotion recognition," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2005, pp. 1468–1471.

[10] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in *Proc. IEEE Int. Conf. on Acoustic, Speech, and Signal Processing*, 2005, vol. 2, pp. 1085–1088.

[11] Z. Zeng *et al.*, "Bimodal HCI-related affect recognition," in *Proc. 6th Int. Conf. Multimodal Interfaces, (ICMI'04)*, 2004, pp. 137–143.

[12] Z. Zeng *et al.*, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 2, pp. 967–972.

[13] G. Bailly, C. Benoit, and T. R. Sawallis, *Talking Machines: Theories, Models, and Designs*. Amsterdam, The Netherlands: Elsevier, 1992.

[14] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. 4th Int. Conf. on Spoken Language Processing*, Philadelphia, PA, Oct. 1996, vol. 3, pp. 1970–1973.

[15] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Classifying emotions in human-machine spoken dialogs," in *Proc. Int. Conf. on Multimedia and Expo*, Switzerland, Aug. 2002, vol. 1, pp. 737–740.

[16] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada, May 2004, vol. 1, pp. 593–596.

[17] T. L. Nwe, F. S. Wei, and L. C. De Silva, "Speech based emotion classification," in *Proc. IEEE Region 10 Conf. Electrical and Electronics Technology*, Singapore, Aug. 2001, vol. 1, pp. 297–301.

[18] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speech signals," in *Proc. Euro. Conf. on Speech Comm. and Tech.*, Geneva, Switzerland, Sept. 2003, pp. 125–128.

[19] T. L. Nwe, F. S. Wei, and L. C. De Silva, "Stress classification using subband features," *IEICE Trans. Inform. Syst.*, vol. E86-D, no. 3, pp. 105–116, 2003.

[20] A. Bartlett, V. Evans, I. Frenkel, C. Hobson, and E. Sumera, Digital Hearing Aids [Online]. Available: www.owlnet.rice.edu/~ elec301/ Projects01/dig_hear_ aid/ 2004

[21] Introduction to Computer Programming With MATLAB Department of Phonetics and Linguistics, Univ. College Landon [Online]. Available: www.phon.ucl.ac.uk/courses/spsci/matlab/, 2004

[22] C. Becchetti and L. P. Ricotti, *Speech Recognition: Theory and C++ Implementation*. Toronto, ON, Canada: Wiley, 1999.

[23] M. J. Lyons, J. Budynek, A. Plante, and S. Akamatsu, "Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis," in *Proc. 4th Int. Conf. Automatic Face and Gesture Recognition*, France, Mar. 2000, pp. 202–207.

[24] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *IEEE Trans. Syst., Man, Cybern, B:Cybern.*, vol. 34, pp. 1588–1595, 2004.

[25] D. Kim and Z. Bien, "Fuzzy neural networks(FNN)-based approach for personalized facial expression recognition with novel feature selection method," in *Proc. IEEE Int. Conf. on Fuzzy Systems*, St. Louis, MO, May 2003, vol. 2, pp. 908–913.

[26] M. Pantic and L. J. M. rothkrantz, "Facial action recognition for facial expression analysis from static face image," *IEEE Trans. Syst., Man, Cybern B:Cybern.*, vol. 34, pp. 1449–1461, 2004.

[27] L. C. De Silva and S. C. Hui, "Real-time facial feature extraction and emotion recognition," in *Proc. 4th Int. Conf. on Information, Communications and Signal Processing*, Singapore, Dec. 2003, vol. 3, pp. 1310–1314.

[28] I. Cohen, N. Sebe, Y. Sun, M. S. Lew, and T. S. Huang, "Evaluation of expression recognition techniques," in *Proc. of Int. Conf. on Image and Video Retrieval*, Urbana, IL, Jul. 2003, pp. 184–195.

[29] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, no. 3, pp. 264–277, Sept. 1999.

[30] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 837–842, Aug. 1996.

[31] S. Theodorids and K. Koutroumbas, *Pattern Recognition*, 2nd ed. New York: Elsevier, 2003.

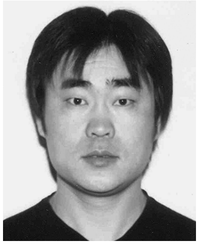
[32] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddale River, NJ: Prentice-Hall, 1999.

[33] J. Ghosh, "Multiclassifier systems: Back to the future, in multiple classifier systems," in Lecture Notes in Computer Science 2002, vol. 2364, pp. 1–15.

[34] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 993–1000, 1990.

[35] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 226–239, 1998.

[36] T. G. Dietterich, "Ensembel methods in machine learning," in *Multiple Classifier Systems*, J. Kittler and T. Roli, Eds. Berlin, Germany: Springer, 2001, vol. 1857, LNCS, pp. 1–15.



Yongjin Wang (S'04) received the M. A. Sc degree in electrical and computer engineering from Ryerson University, Toronto, ON, Canada, in 2005. He is currently pursuing the Ph.D. degree in the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto.

From January to August 2005, he was a Research Assistant at Ryerson Multimedia Research Lab. His research interests include speech and image processing, computer vision, biometrics, and pattern recognition.

Ling Guan (S'88–M'90–SM'96–F'08) received the B.Sc. degree in electronic engineering from Tianjun University, China, the M.A.Sc degree in systems design engineering from University of Waterloo, Waterloo, ON, Canada, and Ph.D. degree in electrical engineering from the University of British Columbia, Vancouver, BC, Canada.

He is currently a Professor and a Tier I Canada Research Chair in the Department of Electrical and Computer Engineering at Ryerson University, Toronto, ON. He also held visiting positions at British Telecom (1994), Tokyo Institute of Technology (1999), Princeton University (2000) and Microsoft Research Asia (2002). He has published extensively in multimedia processing and communications, human-centered computing, pattern analysis and machine intelligence, and adaptive image and signal processing.

Dr. Guan is a recipient of 2005 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS Best Paper Award.