NOTE TO USERS

This reproduction is the best copy available.

RECOGNIZING HUMAN EMOTIONAL STATE FROM AUDIOVISUAL SIGNALS

by

Yongjin Wang
B.Eng., Xi'an Institute of Posts and Telecommunications,
China 1996

A thesis
presented to Ryerson University

in partial fulfillment of the
requirement for the degree of
Master of Applied Science
in the Program of
Electrical and Computer Engineering

Toronto, Ontario, Canada, 2005

Author's Declaration

I hereby declare that I am the sole author of this thesis.
I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

(Yongjin Wang)

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

(Yongjin Wang)

Borrower's Page

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Name	Signature	Address	Date

ABSTRACT

Recognizing Human Emotional State from Audiovisual signals

©Yongjin Wang 2005

Master of Applied Science
Department of Electrical and Computer Engineering
Ryerson University

In this work, we investigate recognition of human emotional state from audiovisual signals. We extract prosodic, Mel-frequency Cepstral Coefficient (MFCC), and formant frequency features to represent the audio characteristic 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 stepwise method based on Mahalanobis distance. The selected features are used to classify the emotional data into their corresponding classes. Different classification algorithms including Gaussian Mixture Model (GMM), K-nearest Neighbors (K-NN), Neural Network (NN), and Fisher's Linear Discriminant Analysis (FLDA) are compared in this study. An adaptive multi-classifier scheme involving the analysis of individual class and combinations of different classes is proposed. Our recognition system is tested over a language independent database. The proposed FLDA-based multi-classifier scheme achieves the best overall and individual class recognition accuracy.

Acknowledgement

I would like to thank my supervisor Dr. Ling Guan for his continuous guidance and encouragement through the course of my research work. I would also like to thank my co-supervisor Dr. Hau San Wong for his enlightening comments and suggestions. This work would have been impossible without their feedback, patience and kindness.

I would also like to thank the Department of Electrical and Computer Engineering for providing a very well equipped and technically supported Ryerson Multimedia Research Laboratory (RML). My thanks are due to the School of Graduate Studies of Ryerson University for providing Graduate Student Scholarship.

I would like to acknowledge my supervisor's funding resource Canada Research Chair program and Canada Foundation for Innovation (CFI) for providing me financial support throughout my research work.

Special thanks are due to the members of Ryerson Multimedia Research Laboratory for their help in collecting the emotional database. It is enjoyable and a great pleasure to work in such a friendly and collaborative environment

In the end, I would like to say thanks to my parents, sisters, and especially, my wife. I could not have achieved this goal without their love, tolerance, encouragement and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General background

AS computers have become an integral part of our lives, 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.

HCI has been critical research area in the multimedia community. It plays critical role in areas such as media indexing and retrieval, media manipulation, design of bio-computers, security and surveillance, and learning of special needs. Components of HCI include face recognition, speech recognition, emotion recognition, gesture recognition, body modeling and movement recognition. A complete HCI system integrates all the aspects of these components.

The main characteristics of human communication are the multiplicity and multimodality of communication channels. Examples of human communication channels are auditory channels which carry speech and vocal intonation, and visual channels which carry facial expressions and gestures. A modality is a sense that is used to perceive signals. Examples of modalities include the senses of sight, hearing, and touch. In real life face-to-face communication, many different channels and modalities are activated, and thus the communication is very flexible and robust. This is also how an

intelligent HCI system should be developed for facilitating robust, natural, efficient, and effective human machine interaction.

In the field of HCI, speech and facial expression are primaries to the objectives of an emotion recognition system. They 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, and educational software [1].

Emotions have been the study of intense interest in both Eastern and Western philosophy since before the time of Lao-Tzu (sixth century B.C.) in the east and of Socrates (470-399 B.C.) in the west, and the most contemporary thinking about emotions in psychology can be linked to one Western philosophical tradition or another [2]. However, the beginning of modern, scientific inquiry into the nature of emotion is thought by many to have begun with Charles Darwin's study of emotional expression in animals and humans [3].

Research into human emotion in psychology and neurophysiology has expanded rapidly in recent years. This fast development is in part due to the advances in imaging technology, but also to a newfound interest in the idea that certain individual emotions may be served by separate brain systems. This latter theoretical position is at the heart of the idea that selected sets of so-called "basic" emotions constitute the foundations of human emotion. All emotions are processed by a circuit of interconnected brain structures known as the limbic system (Figure 1.1 [4]).

Contemporary research in emotion in psychology reveals that certain emotions were associated with distinct facial signals, and that these were common to cultures throughout the world [5]. These basic emotions may be coded by partially distinct brain systems. For instance, the amygdala has been revealed to be playing a significant role in recognizing facial and vocal expressions of fear [6]. A survey of research in psychology about defining, studying, and explaining emotion can be found in [7].

A wide investigation on the dimensions of emotions reveals that at least six emo-

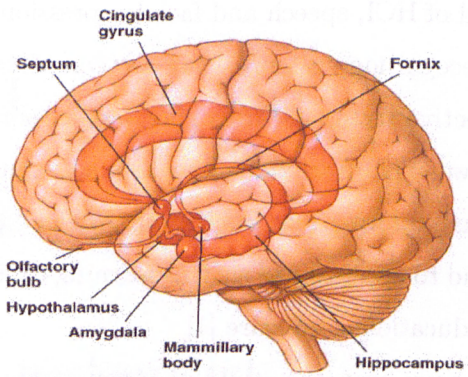


Figure 1.1: The limbic system structure

tions are universal [8]. 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 thesis.

1.2 Motivations

To enhance the human computer intelligent interaction, much effort has been put into machine recognition of human emotions in the past few years. The majority of these works either focus on speech alone, or facial expression only. Relatively few existing works combine these two modalities into one single system for emotion analysis. However, as shown in [9], some of the emotions are audio dominant, while the others are visual dominant. When one modality is failed or not good enough to determine a certain emotion, the other modality can help to perform the recognition. The integration 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.

Due to the fact that most of the existing systems treat audio and visual information separately, no standard database is available for audiovisual recognition of

human emotion yet. Most of the works in the research community utilize a database which they collected on their own. Most of the systems are restricted to a database of only one language. However, the way people convey emotional state might be different according to their cultural background, language, and accent. An efficient emotion recognition system must be able to adapt itself to these aspects. We believe that there are inherent features which are independent of these conditions. The goal of our research is to investigate the independency of these aspects and develop a more generic system to recognize human emotion.

In this thesis, we propose an emotion recognition system to recognize human affective state from audiovisual information. The proposed system is tested over a language, speaker, and context independent database. Audio and visual features are extracted from the emotional data separately to represent the vocal and facial characteristics of humans at different emotions. We perform feature selection to find out the significant features, whilst reducing the dimensionality of the feature space. To achieve higher recognition accuracy, we compare the performance of different classification algorithms. Furthermore, as different emotions might have different significant features, and the features to distinguish combinations of different emotions also might be different, we propose a multi-classifier scheme to analyze these scenarios.

1.3 Data acquisition

In order to conduct the research of recognizing human emotion, a video database that can truly convey the emotional state of human is needed. To our knowledge, there is no such database available. We set up an emotional audiovisual data collecting system to record the emotional video data first.

The performance of an emotion recognition system is highly dependent on the quality of emotional database on which it is trained. When working with speech and facial expression, special care must be taken to ensure that the particular emotion being vocalized and expressed is correct. We set up a data collecting system to record emotional data with adequately high accuracy. A digital video camera was used to

record the samples in a 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 happening, 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. Examples of these sentences are as follows: I am so happy today (Happiness); Why do you always cheat on me (Anger)? I messed up the mid-term (Sadness); What? Is that true (Surprise)? Please, please do not beat me (Fear); That bad fish has a disgusting smell (Disgust). 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 for each emotional class, 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 cultural background, language, and accent. 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. Some of the subjects have facial hair, which further increase the diversity of the database. We collected a total of 500 video samples, each delivered with one of the six particular emotions. The samples were recorded at a sampling rate of 22050 Hz, using a single channel 16-bit digitization. Figure 1.2 shows examples of the emotional data.

1.4 Outline of thesis

The remainder of this thesis consists of 6 chapters which are organized as follows:

Chapter 2: *Literature Review*, reviews the related works reported in the literatures. Recognizing human emotion from audio, video, and audiovisual information are reviewed separately.

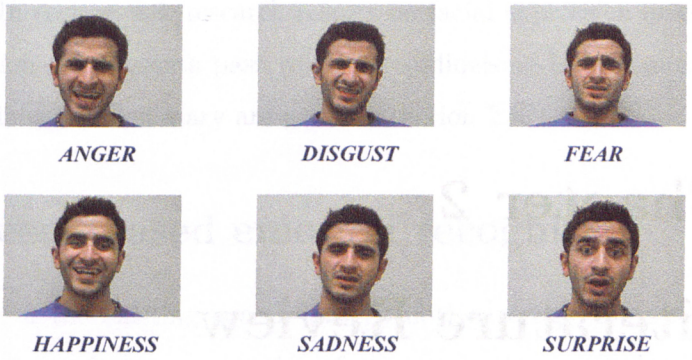


Figure 1.2: Video clips of facial expressions portrayed by one subject

Chapter 3: *Audio Feature Extraction*, discusses the audio features that we extracted to represent the vocal characteristic of human speech at different emotions.

Chapter 4: *Visual Feature Extraction*, describes the HSV color model based face detection scheme, and Gabor wavelet features that we extracted for facial expression representation.

Chapter 5: *Feature Selection and Classification*, details the applied feature selection and classification algorithms.

Chapter 6: *Emotion Recognition System*, presents and compares the experimental results of different classification and feature selection algorithms. An adaptive multi-classifier scheme is introduced to enhance the performance of the system.

Chapter 7: *Conclusions*, discusses the results and summarizes the contributions of this work. Some considerations are provided on how to improve this work in the future.

Chapter 2

Literature Review

2.1 Introduction

UNDERSTANDING human emotional state has been a great challenge which involves multi-discipline research from psychology to signal processing. Recent advances in speech/image analysis and pattern recognition open up the possibility of automatic detection and classification of emotional audiovisual signals. A great deal of studies has been conducted in the research community. Applications of emotion recognition can be foreseen in the broad area of human computer intelligent interaction.

Emotion analysis is greatly influenced by inputs such as voice, facial expression, body language, and understanding of the semantic meaning of the words, etc. However, it is widely recognized that speech and facial expression are the two major modalities in human communication, and thus for most of the human-computer interaction applications. The body language that humans express in different emotions is highly dependent on cultural background, personality, gender, age, etc. It is hard to model the emotional behavior of an individual in a general way. Although understanding the semantic meaning of the words in a certain language might help to recognize emotion, two sentences could have the same lexical meaning but different emotional information. So far, the majority of studies utilize either speech or facial expression separately. Yet a few systems combine both of these two modalities.

In this chapter, we will first review the recent works on speech-based emotion

recognition. In section 2.3, research results on facial expression analysis will be discussed. Section 2.4 presents past works in audiovisual-based emotion recognition, while discussions and summary are given in section 2.5.

2.2 Speech-based emotion recognition

The majority of the existing speech-based emotion recognition systems consists of three components: feature extraction, feature selection, and classification (Figure 2.1). The major challenge here is to extract features that can truly represent the vocal characteristics of human speech at different emotional states. There are basically two types of features: prosodic features and phonetic features. Prosodic features are mainly related to the rhythmic aspects of the speech, while phonetic features focus on the linguistic aspects of the speech. After speech feature have been extracted, feature selection algorithms are applied to select significant features, while classifiers are used to map the features to the corresponding emotional state. Table 2.1 provides a summary of the methods that were proposed in some of the past works.

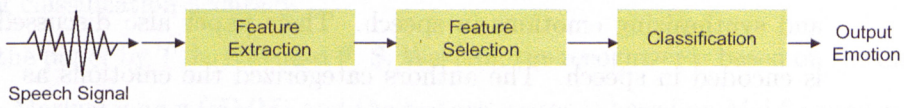


Figure 2.1: Architecture of speech-based recognition system

Cowie and Douglas-Cowie [10] broadened the view of emotion recognition by noting that speech features previously associated with emotion could also be associated with impairments such as schizophrenia and deafness. Their studies focussed more on intensity and rhythm, and also investigate the importance of unvoiced sounds.

Amir and Ron [11] reported a method for identifying the emotions based on analysis of the signal over sliding windows. A set of basic emotions was defined and for each emotion a reference point was computed. At each instant the distance of the measured parameter set from the reference point was calculated, and used to compute a fuzzy membership index for each emotion.

Reference	Features Extraction	Feature Selection	Classification Algorithms
Dellaert [13]	Statistics of: Rhythm, Pitch, Voiced parts, Slopes	PFS, FS, Voting	Maximum Likelihood, Kernel Regression, K-NN
Nicholson [14]	Power, Pitch, LPC, Delta LPC		Neural Network
Lee [15]	Statistics of: Pitch, Energy	PCA	Linear DA, K-NN, SVM
Nwe [16]	Mel frequency power coefficients		HMM
Kwon [17]	Pitch, Energy, MFCCs, Formant	FS/BE	SVM, DA, HMM
Schuller [18]	Pitch, Energy, HMM-based ASR engine		MLP Neural Networks
ververidis [19]	Statistics of: Pitch, Energy, Spectrum	FS, PCA	Nearest Mean, Bayes Classifier

Table 2.1: Summary table of speech-based recognition

Sato and Morishima [12] demonstrated the use of neural networks in analyzing and synthesizing emotions in speech. Their paper also discussed the way emotion is encoded in speech. The authors categorized the emotions as "anger", "disgust", "happiness", "sadness", and "neutral". They briefly discussed the six speech parameters that were analyzed, and made some empirical observations about the relationship between these parameters and certain emotions. The sound samples used in the experiment were recorded by a professional actor, yet a group of human subjects could correctly recognize his emotion only 70% of the time.

Dellaert et al [13] explored several statistical pattern recognition techniques to classify utterances according to their emotional content. Different feature selection methods including promising first selection (PFS), forward selection (FS), and majority voting of specialist are compared to improve the system performance. A set of four emotions: happy, sad, anger, and fear, is used as the emotion categories. The applied emotional speech database contains 1000 speech samples from 5 human sub-

jects. This system achieves a recognition rate of up to 79.5%. However, the authors noted in their concluding remarks that the results were speaker dependent, and that the classification methods had to be validated on completely held-out databases.

Nicholson et al [14] proposed a speaker and context independent system for emotion recognition in speech using neural networks. The paper examined both prosodic features and phonetic features. A set of eight emotions, namely joy, teasing, fear, sadness, disgust, anger, surprise, and neutral, is investigated. A large database of totally 100 speakers, 50 male and 50 female, using phoneme balanced words were used to test the system. This database ensures the speaker and context independence of the experiments. However, no feature selection techniques were applied to get the best feature subset, and the recognition rate was only around 50%.

In the paper by C. M. Lee, S. Narayanan, and R. Pieraccini [15], a corpus of human-machine dialogs recorded from a commercial application is classified into two basic emotions: negative and non-negative. Three methods are used to classify the emotional utterances. The features used by the classifiers are utterance-level statistics of the fundamental frequency and energy of the speech signal. Principal component analysis (PCA) was applied to reduce the dimensionality of the features whilst maximizing classification accuracy.

In the paper by T. L. Nwe and F. S. Wei [16], the recognizer is based on the discrete hidden Markov model (HMM) and the feature vector is based on Mel frequency short time speech power coefficients. Six basic emotions: anger, dislike, fear, happiness, sadness, and surprise are investigated. A universal code book is constructed based on emotions under observation for each experiment. The database consists of 90 emotional utterance each from two speakers. The system gives an average accuracy of 72.22% and 60% respectively for the two speakers.

Kwon et al [17] select a group of prosodic and phonetic features as the input to a forward selection (FS) and backward elimination (BE) method used for feature selection. Different classification algorithms including support vector machine (SVM), discriminant analysis (DA), and HMM are compared. However, this system only achieves 42.3% for 5-class emotion recognition.

Schuller et al [18] introduce an approach to the combination of acoustic features and language information for automatic emotion recognition in an automotive environment. The extracted acoustic features are represented by the statistics of pitch and energy, while the language model is a standard hidden Markov model based automatic speech recognition (ASR) model. A multi-layer perception (MLP) neural network is engaged to fuse the two types of speech features.

Ververidis et al [19] extract 87 speech features from a corpus of 500 emotional utterance collected from four subjects. Sequential forward selection is used to discover significant features with cross-validation as the criterion. Principal component analysis is then applied to the selected features to further reduce the dimensionality. Two classifiers are used to classify the utterances into five classes. This system only achieves an overall accuracy of 51.6%. The authors also provides clues that the system can be improved by reducing classification to a two-class problem, due to the fact that the features which can separate two classes could be different from those which separate five classes.

2.3 Facial expression analysis

To perform human emotion recognition from facial image or image sequence, the face region should be detected from the image first. Then facial expression information can be extracted from the observed facial image or image sequence. In the case of still image, extracting facial expression information means to localize the face and its features in the image. In the case of facial image sequence, it means to track the motion of the face and its features in the image sequence. The facial features are usually the prominent components of the face, such as mouth, eyes, nose, eyebrow, and chin. However, we might also use face model features to represent the face. This kind of model can represent the face as a whole unit (holistic representation), as a set of features (analytic representation), or as a combination of these two (hybrid approach). The applied face representation and the kind of input images determine the choice of mechanisms for facial expression information extraction [20].

After the face and its appearance have been perceived, the next step is to identify the facial expression conveyed by the face. This involves classifying facial expressions into a set of predefined categories that we want to deal with. Although the categorization of human emotion is a rather complicated issue, most of the studies on facial expression analysis utilize Ekman's emotional categorization of facial expression [5], in which six basic emotions, namely happiness, sadness, surprise, fear, anger, and disgust, are defined. Figure 2.2 depicts the architecture of a facial expression analysis system.

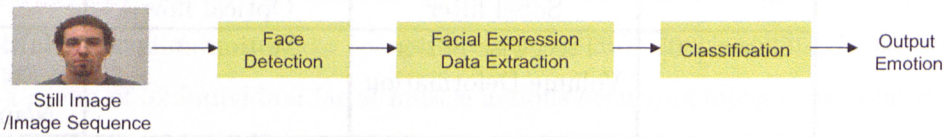


Figure 2.2: Architecture of facial expression analysis system

Facial Action Coding System (FACS) proposed by P. Ekman and W. V. Friesen [20] is one of the most popular representations of facial expression. It is based on the enumeration of all action units (AU) of a face that cause facial movements. There are totally 46 AUs in FACS that account for changes in facial expression. The combination of these action units result in a large set of possible facial expression. A trained human FACS coder decomposes an observed expression into specific AUs that produced the expression. FACS is coded from video and the code provides precise specification of the dynamics (duration, onset and offset time) of facial movement in addition to the morphology (the specific facial actions which occur). This FACS model is widely used in facial expression recognition research [21][22][23]. By tracking facial features and measuring the amount of facial movement, different facial expression can be categorized.

Ekman's work inspired many computer vision researchers to perform facial expression analysis by means of image and video processing. Samal and Iyengar [24] provides an overview of the early works in 1992, while Pantic and Rothkrantz give a comprehensive review of the works in the late 1990s [25]. In this section, we will

review some of the recent works in facial expression analysis. Table 2.2 summarize the methods that are adopted in some of the papers.

Reference	Face Detection	Feature Extraction	Classification Algorithms
Haneda [26]	HSV color model, Movement	Eye, Mouth	Expert System,
Lyons [27]		2-D Gabor wavelet	LDA
Pantic [28]	Watershed, HSV color model	Profile contour, Contour of facial components	Rule-based reasoning
Silva [29]	Low-pass filtering, Sobel filter	Edge counting, Optical flow	Neural Network
Cohen [30]	Piecewise bezier Volume Deformation	Motion units	Multi-level HMM, NB, TAN, Neural Network
Ma [31]		2-D DCT	Constructive Neural Network
Kim [32]	Color histogram in PCA-based color space	Eye, Eyebrow, Mouth, Wrinkles, Furrows	Fuzzy Neural Network

Table 2.2: Summary table of facial expression analysis

Haneda et al [26] use expert system to classify facial expression. The facial area is extracted based on the Hue and Quasi-chroma element images in the modified HSV color system and the movement information of the object in the video frames. The extracted facial area is then normalized using the center points of the eyes, the tip of the nose and two corner points of the mouth. Feature amounts of the eye, mouth and other areas of the face are then extracted using different methods. The automatic generation of knowledge is achieved by automatic construction of membership functions. Based on the membership functions constructed and by combining all the information, the rule sets for recognition of expression are constructed. This system is tested on a total of 4410 video frames of three people and achieves an accuracy of more than 90% on average.

Lyons et al [27] use a set of multi-scale, multi-orientation Gabor filters to transform the images first. A grid is then automatically registered with the face using a elastic

graph matching method. The Gabor coefficients sampled on the grid are combined into one single vector as the features. Principal Components Analysis (PCA) is applied to reduce the dimensionality of the feature space. They test their system with a database of 193 images posed by 9 Japanese females, and achieve 75% expression classification accuracy by using Linear Discriminant Analysis (LDA) .

Pantic and Rothkrantz [28] propose an automatic facial action recognition system using dual-view static image. The face is detected using a watershed segmentation with markers method, in which the markers are extracted based on HSV color model. A multi-detector approach to facial feature localization is utilized to spatially sample the profile contour and the contours of the facial components such as the eyes and mouth.They report an average recognition rate of 86% by classifying facial actions into a group of 32 individual facial muscle actions occurring along or in combination using rule-based reasoning.

Silva and Hui [29] determine the eye and lip position using low-pass filtering and edge detection methods. Edge counting and image-correlation optical flow techniques are applied to calculate local motion vectors of facial features. They achieved an average emotion recognition rate of 60% using a neural network.

Cohen et al [30] introduce and test different classifiers for recognizing human facial expression from video sequences. A face tracking algorithm called Piecewise Bezier Volume Deformation tracker (PBVD) is used, and 12 Motion-Units (MU) are extracted as the basic features for classification. They introduce a multi-level HMM classifier for dynamic classification, in which the temporal information is also taken into account. Two types of Bayesian network classifier, Naive Bayes (NB), and Tree-Augmented Naive Bayes (TAN), and neural network are investigated to perform classification on a single frame. Person-independent experiment using their own database shows that TAN classifier gives the best correct recognition rate of 66.53%.

Ma and Khorasani [31] use two-dimensional (2-D) discrete cosine transform (DCT) over the entire face image as a feature detector, and a constructive one-hidden-layer feed-forward neural network as a facial expression classifier. The employed image

database contains face images that only have face in the scene. Experimental results show that the constructive neural network outperforms vector matching method and fixed-size neural network.

Kim and Bien [32] use a color histogram based adaptive threshold method in PCA-based color space to detect the face region. Five features are extracted from the facial components to represent the facial expression, where the facial components are extracted using a course-to-fine approach. They also applied a histogram-based algorithm to select features. This system achieves 94.3% accuracy by using a fuzzy neural network for classification.

2.4 Bimodal emotion recognition

Although speech-based emotion recognition and facial expression analysis have been studied extensively in the past, audiovisual based emotion recognition has been rarely investigated in the literatures. Silva et al. [9] report the results of human subjects' perception of human emotions. They found that some emotions (sad,fear) are strongly auditory dominant, some emotions (happy, surprise and anger) are strongly visually dominant, while some (disgust) are mixed dominant. This implies that the integration of audio and visual information will potentially outperform any of them alone.

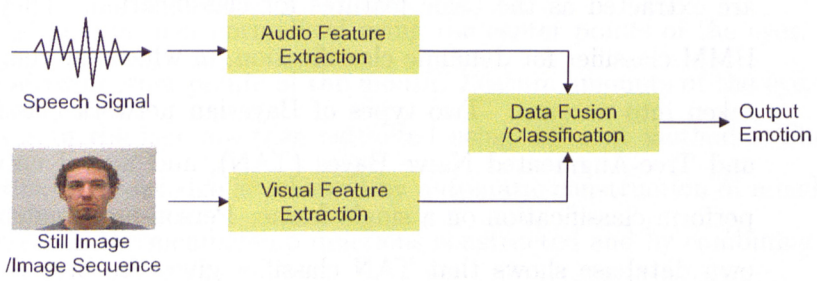


Figure 2.3: Architecture of bimodal emotion recognition

The architecture of an audiovisual emotion recognition system is depicted in Figure 2.3. Audio and visual features are extracted from speech and images respectively. These two streams of data are then fused into a classifier for emotion recognition. In

this section, we review the existing audiovisual based emotion recognition systems, with the summary of the applied methods in Table 2.3.

Reference	Audio Features	Video Features	Data Fusion /Classification
Song [33]	Pitch, Energy	Eyebrow, Eyelid, Cheek, Lips, Jaw	Triple HMM
Silva [34]	Pitch	Lips, Eyebrows, Mouth corners	Nearest Neighbors HMM, Rule-based method
Chen [35]	Pitch contour, Energy envelope	Eyebrow, Cheek, Mouth, Facial components	Nearest Mean, Gaussian Model
Go [36]	Wavelet coefficient	DWT-based fisherface	Multi-decision scheme PCA, Rule-based method

Table 2.3: Summary table of bimodal emotion recognition system

Song et al [33] propose an audiovisual emotion recognition system based on hidden Markov model. The extracted audio features are pitch and energy features. The motion of eyebrow, eyelid, and cheek are extracted as expression features, while that of lips and jaw as the visual speech ones. The extracted three-stream audiovisual features are fused into a triple HMM for classification. This system was tested by a database of 684 samples, and an overall accuracy of 85% was claimed.

Silva and Ng [34] build an audio and a video system separately. In the audio system, pitch is extracted as the features and a nearest neighbors method is used for classification. In the video system, They track the edge movement of lips, mouth corners, and eyebrows using an optical flow algorithm, while hidden Markov model is trained as the classifier. A rule-based system is adopted to fuse the results of audio and video classification. The recognition results show that the bimodal approach outperforms the individual systems.

Chen et al [35] utilize statistics of the pitch contour, energy envelope and their derivatives to represent the characteristics of emotional speech, while that of the visual information is obtained by tracking the position of eyebrow, cheek lifting, and mouth opening. The audio and video features are simply concatenated into one vector for classification. They compared two methods for classification: nearest-mean,

and Gaussian model. By using leave-one-out cross validation, the bimodal approach achieves 97.2%, which greatly improved the recognition rate from audio-only, 75%, and video-only, 69.4%.

Go et al [36] also demonstrate the advantage of bimodal emotion recognition over single modality. In their system, they apply principal component analysis (PCA) to classify the facial expression features, which are extracted using a multi-resolution analysis based on the discrete wavelet transform (DWT). For emotion recognition from speech signal, a multi-decision making scheme is proposed to perform classification independently on each wavelet subband. Finally, the facial and speech recognition methods are merged using a rule-based scheme, and the better performance is achieved.

2.5 Discussions and summary

In this chapter, we have reviewed and compared some of the works reported in the literature in the research area of human emotion recognition. The audiovisual based approach has been demonstrated to perform better than single modality based methods [33][34][35][36] due to the fact that more information is conveyed and the complementary relationship of these two modalities helps. Overall, a bimodal emotion recognition system is composed of a few components: audio feature extraction, visual feature extraction, feature selection, and classification. In this section, we will give remarks on these aspects separately.

Audio feature extraction

In the field of audio processing, there are basically two types of features: prosodic feature and phonetic feature. Prosodic features deal with the more musical aspects of speech, such as rising or falling tones and accents or stresses. Phonetic features deal with the types of sounds involved in speech, such as vowels and consonants and their pronunciation. Speech understanding traditionally uses only phonetic features. In speech-based human emotion recognition, most of the works focus on prosodic features

only [13][15][19][33][34][35], which can be represented as the statistic features of pitch, energy, slope, and speaking rate, etc. However, as shown in [14][16][17][18], phonetic features, which represent the linguistic factor of the speech, might also contribute to emotion recognition, and thus should also be considered.

A major restriction of the existing systems is that, the scope of their research has been limited to one language. However, the way people express their emotional state in speech might be different according to their cultural background, language, accent, etc. For a more generic application, an emotion recognition system should be capable of recognizing human affective state regardless of these aspects. This is one of the major goals of our research.

Visual feature extraction

The successful and accurate detection of human face from the scene can greatly reduce the noise and negative affects of the background, and thus is very critical in facial expression analysis. The adoption of face detection scheme is highly dependent on the type of input images [37]. The majority of the past works utilize either edge [29] or color [26][28] information for face detection. In certain human computer interaction applications, if the face is frontal viewed and the background is not very complex, color information is an efficient tool for identifying facial areas and specific facial features. The color based segmentation technique has the advantage of simple and fast, which is very important for real life application.

After the facial area being detected, human emotion can be recognized by investigating the movements of points belonging to the facial features such as mouth, eye, eyebrow and then analyzing the relationships between those movements [22][30][33][35]. This image sequence based technique might achieve higher accuracy, but the computational complexity is high. Extracting a key frame from the image sequence to represent the characteristics of the whole video is an option to achieve tradeoff between accuracy and computational complexity. The analysis of emotion recognition from static images [26][27][28] also demonstrate the efficiency of still image based recognition.

Regarding facial feature extraction from still images, both holistic [27] and analytic [26] approaches have been studied in the past. The holistic approach treat the human face as a whole unit, while the analytic approach tries to analyze the prominent features of the human face, such as eyes, eyebrows, and mouth. Although human facial expression can be reflected by a few components of human face in a certain extent, some of the information is also lost. A holistic approach will potentially better capture the characteristics of human face at different emotions.

Feature Selection

Feature selection is a very important step in pattern recognition problems. Usually, a large number of features are extracted from the data source. But making complete use of these features might not achieve the best results. A large features space also cause the problem of curse of dimensionality, and the system becomes inapplicable. How to reduce the size of feature space without losing too much information and get better results inspires the research in feature selection.

Different dimensionality reduction and feature selection algorithms have been studied in the past [38][39]. The criteria to select a feature selection algorithm is to get a compromise between efficiency and computational complexity. Suboptimal methods such as forward selection and backward elimination is a good tradeoff between these two aspects, and thus are frequently used in the field[13][17][19].

Classification

Classification algorithms can be categorized into two groups: supervised and unsupervised, while supervised method can be further divided into linear and nonlinear classifiers. The selection of classification algorithm depends on the problem on hand. For human emotion recognition, we are trying to classify human emotion into a few predefined basic emotions by training the machine through class-known samples. It is a supervised case.

Usually, we do not have complete knowledge of the data that will be applied to the classifiers, and thus different classification algorithms need to be compared. Different

algorithms including k-nearest neighbors, discriminant analysis, neural network, hidden Markov model, support vector machine, Bayes classifier, have been used in the literature. In audiovisual based emotion recognition, some papers utilize rule-based scheme to perform fusion of audio and visual classification [34][36]. This rule-based method usually based on priori knowledge of the audio and video factors in the system.

Most of the past works applied one classifier to classify multi-class problems. However, the significant features selected in a two-class problem might not be the same as that of in a multi-class scenario [19]. Individual class also has different features to distinguish from other classes. In this thesis, we compare the performance of some classification algorithms first. Based on the experimental results, we introduce a multi-classifier scheme to analyze these scenarios. This divide and conquer method helps to achieve higher recognition accuracy.

Chapter 3

Audio Feature Extraction

3.1 Introduction

HUMAN are capable of detecting other human's emotions by listening to their voices. Although different languages and accents are used worldwide, and the way people express their emotions in speech varies according to their cultural background, personality, age, gender, etc., in most of the cases, we can perceive other peoples's feelings. For more natural HCI applications, we need to give computers the same capability as human does. As one of the major indicators of human affective state, speech plays an important role in machine recognition of human emotion.

To build a more generic emotion recognition system, the extraction of features that can truly represent the universal characteristics of the intended emotion is a real challenge. 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 [40], most of the researches adopt prosodic features. However, Mel-frequency cepstral coefficient (MFCC) and formant frequency are also widely used in speech recognition and some of the other speech processing applications. 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 also investigate these two types of features.

Figure 3.1 shows the block diagram of the audio feature extraction system used in this thesis. It consists of five components. The preprocessing step performs noise

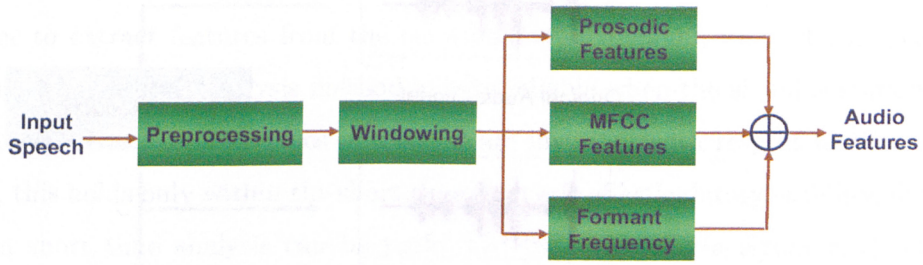


Figure 3.1: Audio feature extraction system

reduction and silence elimination. Then the preprocessed signal is passed through a windowing process to segment the original signals into short time speech frames. Prosodic, MFCC, and formant frequency features are then extracted separately based on short time spectral analysis. In this chapter, we will discuss each of the components in detail.

3.2 Preprocessing

The emotional data that we used in this thesis is recorded in a relatively quiet environment. However, the "hiss" of the recording machine and background noise can greatly effect the extraction of features from the speech signal, and thus the performance of the recognition system. Furthermore, the leading and trailing edge produced in the recording process do not provide any useful information, and are totally redundant. To reduce the effects of noise and silence in the speech utterance, we perform noise reduction and leading and trailing edge elimination at the preprocessing stage. Figure 3.2 shows the flow of preprocessing.

In this work, the "hiss" of the recording machine and background noise are reduced by thresholding the wavelet coefficients [41]. The traditional low-pass filtering methods, which are linear time invariant, can blur the sharp features in a signal and sometimes it is difficult to separate noise from the signal where their Fourier spectra overlap. The wavelet method has the advantage of reducing the noise efficiently

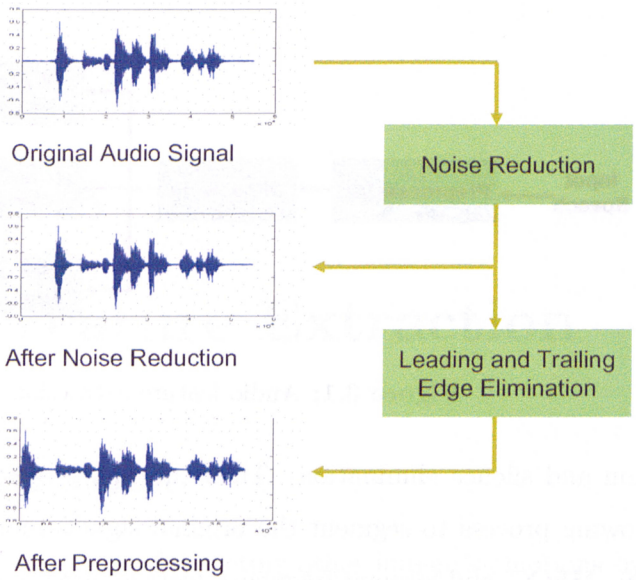


Figure 3.2: Flow of preprocessing stage

without blurring the features in the original signal. For wavelets, which are nonlinear functions, the amplitude, instead of the location, of the Fourier spectra, differs from that of the noise. This allows for thresholding of the wavelet coefficients to remove the noise. If a signal has energy concentrated in a small number of wavelet coefficients, their values will be large in comparison to the noise that has its energy spread over a large number of coefficients. These localizing properties of the wavelet transform allow the filtering of noise from a signal to be very effective. While linear methods trade-off suppression of noise for broadening of the signal features, noise reduction using wavelets allows features in the original signal to remain unchanged.

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. The margin value is estimated based on experiments. We then exclude the leading and trailing edge by eliminating those silent parts with amplitude below the threshold. In this work, the predefined period is 250ms, and a value of 0.01 is selected as the margin value.

3.3 Windowing

In order to extract features from the emotional speech signal, we perform spectral analysis. The spectral analysis method is only reliable when the signal is stationary, i.e. the statistical characteristics of a signal are invariant with respect to time. For speech, this holds only within the short time intervals of articulatory stability, during which a short time analysis can be performed by windowing a signal $x(n)$ into a succession of windowed sequences $x_t(n)$, called frames. These speech frames can then be processed individually.

$$x_t(n) = w(n)x'_t(n) \quad n = 0, ..., N-1, \quad t = 0, ..., T-1, \tag{3.1}$$

where $w(n)$ is the impulse response of the window, N is the size of the window, T is the number of frames, and $x'_t(n)$ is the frame before applying the window function. In this thesis, the adopted window is Hamming window (Figure 3.3), whose impulse response $w(n)$ is a raised cosine impulse:

$$w(n) = 0.54 - 0.46cos(\frac{2\pi n}{N-1}) \quad n = 0, ..., N-1. \tag{3.2}$$

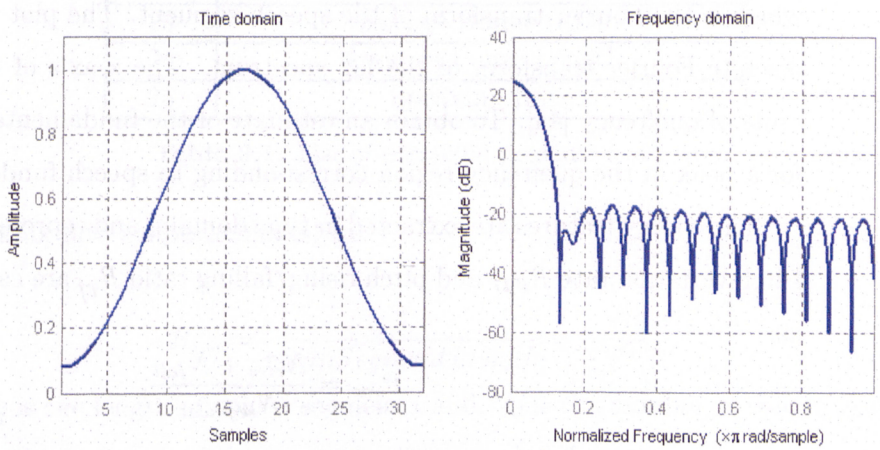


Figure 3.3: Hamming window (32 points). **Left:** time domain, **Right:** frequency domain

Compared with the rectangular window shape, Hamming window has the advantage of decreasing the leakage effect [42]. The size of the window is determined

by considering a tradeoff between time and frequency resolution. Furthermore, to smooth the transition and eliminate the possible gaps between blocks, overlapping windows are usually employed. In this thesis, the spectral analysis is performed over speech frames of 512 points, with 50% overlap.

3.4 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 thesis, we extracted 25 prosodic features as listed in Table 3.1.

The pitch is estimated based on the Fourier analysis of the logarithmic amplitude spectrum of the signal [43][44]. By taking the logarithm of the frequency spectrum, the low frequency components of the spectrum are exaggerated. If the log amplitude spectrum contains many regularly spaced harmonics, then the Fourier analysis of the spectrum will show a peak corresponding to the spacing between the harmonics: i.e. the fundamental frequency. This method is demonstrated in Figure 3.4. The top plot is the waveform of a time domain signal. The plot in the middle shows the spectrum, which is the Fourier transform of the speech segment. The plot at the bottom is the discrete Fourier transform of the log spectrum. The x-axis of the bottom plot has units of quefrency [44]. To obtain an estimate of the fundamental frequency, we look for a peak in the quefrency region corresponding to speech fundamental frequencies.

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:

$$R_{var} = \frac{N_{rise} + N_{fall}}{N_{segment}}, \tag{3.3}$$

$$R_{rf} = \frac{N_{rise}}{N_{fall}}, \tag{3.4}$$

where N_{rise} is the number of pitch rise; N_{fall} is the number of pitch fall; and $N_{segment}$ is the number of segments.

Index	Feature Description
1	Pitch Mean
2	Pitch Median
3	Pitch Standard Deviation
4	Pitch Max
5	Pitch Range
6	Pitch Variation Rate
7	Pitch 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

Table 3.1: List of extracted prosodic features

Speaking rate is approximated by:

$$R_{spk} = \frac{1}{meanSegmentLength} = \frac{N_v}{\sum_i^{N_v} T_i}, \tag{3.5}$$

where T_i is the length of voiced segment i and N_v is the number of voiced segments.

Pitch slope of each of the rise and fall is calculated as:

$$S_{pitch} = \frac{pitchDifference}{lengthOfRiseOrFall} = \frac{|f_{max} - f_{min}|}{t_{end} - t_{start}}, \tag{3.6}$$

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

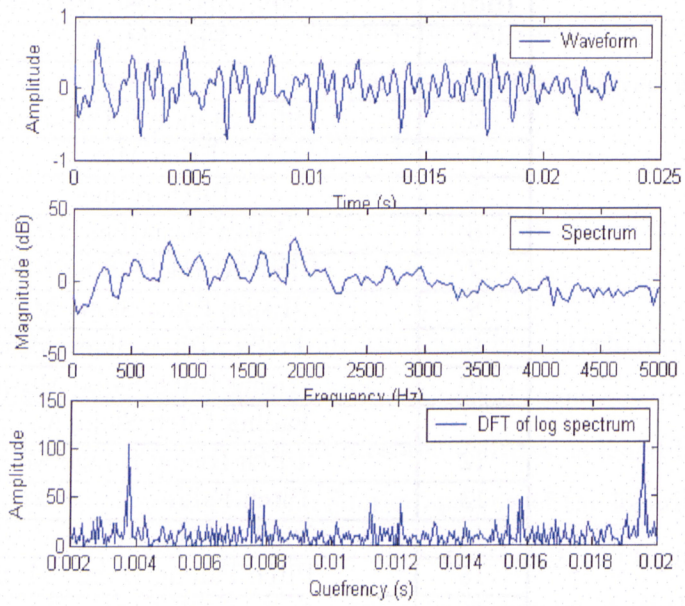


Figure 3.4: Pitch estimation. **Top:** waveform in time domain, **Middle:** magnitude spectrum, **Bottom:** DFT of log magnitude spectrum

Pauses are of no use in calculating the other parameters, so they are discarded. Pitch (amplitude) range is determined by scanning the curve, finding the maximum and minimum pitch (amplitude), and calculating the difference.

Mean values and standard deviation are calculated respectively by:

$$\bar{x} = \sum_{i}^{N} x_i p(x_i), \tag{3.7}$$

$$\sigma = \sqrt{\sum_{i=0}^{N} (x_i - \bar{x})^2 p(x_i)}, \tag{3.8}$$

where N is the range of x_i and $p(x_i)$, the probability of values x_i occurring, is calculated using a normalized histogram. In other words, the number of times x_i occurred in the speech signal is found, then divided by the total number of sampling points. Median values are then calculated by finding the value of x_i such that:

$$\sum_{i} p(x < x_i) \approx 0.5. \tag{3.9}$$

3.5 MFCC features

Mel-frequency cepstral coefficient (MFCC) is perhaps the most popular solution in the field of speech recognition, identification, etc. The purpose of MFCC is to mimic the behavior of human ears by applying cepstal analysis. As our goal is to identify possible acoustic features that can contribute to the recognition of human emotion, we also investigate this type of feature.

Calculating the MFCCs for a speech signal consists of preprocessing, windowing, followed by Fourier transform , Mel-scaling and inverse cosine transform for each time frame. Prior to inverse transform, the magnitude of the spectrum is made logarithmic. This logarithmic scale is a characteristic of the human hearing system. Figure 3.5 shows the flow of the MFCC feature extraction procedure.

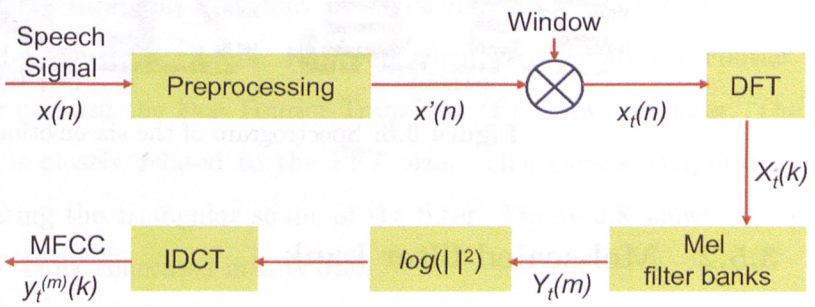


Figure 3.5: MFCC Processing flow

3.5.1 Spectral analysis

The preprocessing and windowing process described in section 3.2 and 3.3 are also applied here. In order to perform spectral analysis, the speech signals need to be transformed to the frequency domain. This is done by the discrete Fourier transform. Figure 3.6 shows spectrograms and associated waveforms of the six emotions, as produced by one of the experimental subjects. On the spectrogram, time is represented along the horizontal axis, whereas frequency is plotted along the vertical axis. For a given spectrogram S, the strength of a given frequency component f at a given time t

in the speech signal is represented by the darkness of the corresponding point $S(t,f)$. It can be observed that each emotion class exhibits different patterns.

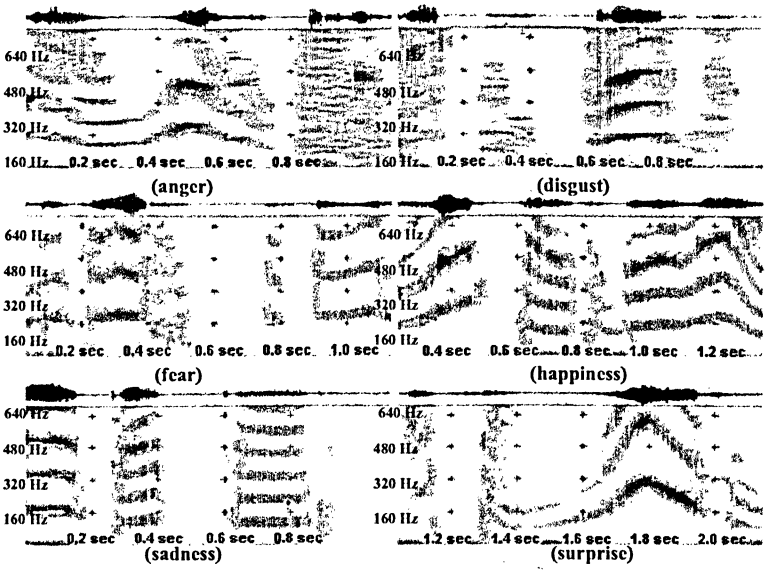


Figure 3.6: Spectrogram of the six emotions

3.5.2 Mel-scaled filter bank

In order to simplify the spectrum without significant loss of data, the Fourier transformed signal is usually passed through a set of band-pass filters, which properly integrate a spectrum at defined frequency ranges. In a speech signal, most of the important and useful information is located at the lower frequency band. Mel-scale is the most widely used perceptual scale, which is designed to capture and emphasize the information in low frequency band. The filter bank is usually constructed of triangular-shaped filters with frequency overlap, so that the center frequency of a filter corresponds to the upper frequency of previous filter and lower frequency of next filter. The central frequency of each Mel filter bank is uniformly spaced before 1 kHz and it follows a logarithmic scale after 1 KHz. Furthermore, to emphasize the low frequency components, the filter magnitude is usually set to 1 at the low frequency band, while decreasing as the frequency increases.

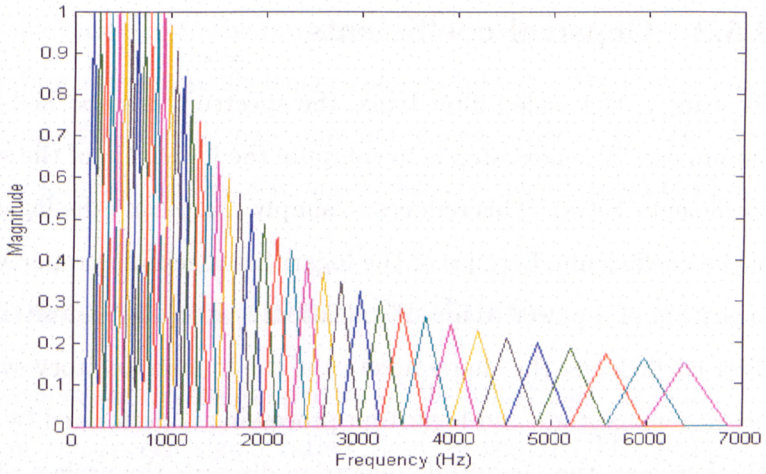


Figure 3.7: Mel-scaled filter bank design

Usually, the range of the frequency covered by the filter bank lies between 20 Hz till half of the sampling frequency of the signal. Figure 3.7 shows the diagram of an ideal Mel-scaled filter bank. However, when we calculate the Fourier transform, we usually perform the Fast Fourier Transform (FFT) by computer. The frequency resolution is closely related to the FFT size. This causes computational error in approximating the triangular shape of the filter. Figure 3.8 shows the arrangement of the filter bank generated in this work.

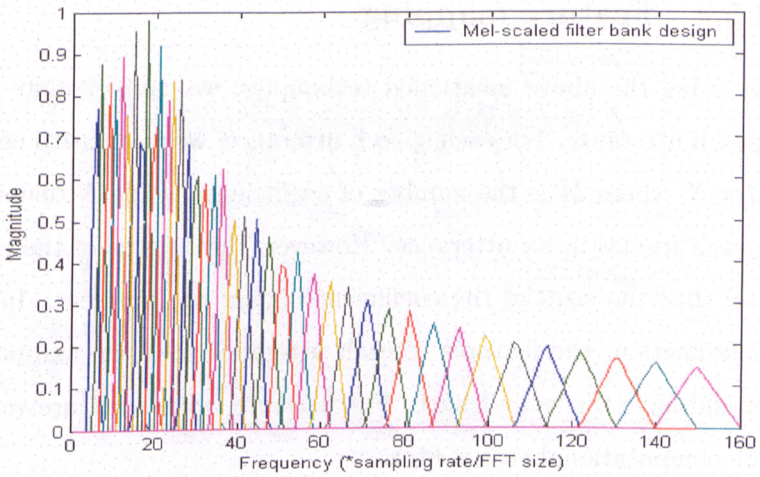


Figure 3.8: Mel-scaled filter bank generated in this work

3.5.3 Cepstral coefficients

By using a Mel-scaled filter bank, the spectrum is smoothed in a way similar as the human ear. The next step is to compute the logarithm of the square magnitude of the coefficients $Y_t(m)$. This reduces to simply computing the logarithm of the magnitude of the coefficients, because of the logarithm algebraic property which brings back the logarithm of a power to a multiplication by a scaling factor. By taking the log of the filter coefficients, the characteristics of the human auditory system can be simulated, because magnitude and logarithm processing is performed by the human ear as well. Furthermore, the magnitude operation discards the useless phase information, while a logarithm performs a dynamic compression, making feature extraction less sensitive to variations in dynamics [42].

MFCCs are the inverse discrete cosine transform of the logarithm of the magnitude of the filter bank output:

$$y_t^{(m)}(k) = \sum_{m=1}^{M} \log\{|Y_t(m)|\} \cdot \cos(k(m - \frac{1}{2})\frac{\pi}{M}), \tag{3.10}$$

where m is the index of the filter, and M is the total number of the filters. $y_t^{(m)}(k)$, $m=1,2,...,M$ represent the Mel-frequency cepstral coefficients.

3.5.4 Feature mapping

By using the above mentioned techniques, we can calculate the MFCCs from each speech utterance. For each speech utterance, we calculate a coefficients matrix of size $M \times N$, where M is the number of coefficients, while N denotes the total number of speech frames in an utterance. However, the lengths of the utterances are different, and thus the sizes of the coefficient matrix are different. In order to facilitate the classification, the features of each utterance that are 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 transformation. In this work, we take the first thirteen coefficients as the useful data. We then calculate the mean, median, standard deviation, max, and min of each order of $y_t^{(m)}(k)$, $m=1,2,...,13$ as the extracted features. In this way, we extracted a total number of 65 MFCC features.

3.6 Formant frequency features

Formant frequencies are the properties of the vocal tract system. It needs to be inferred from the speech signal rather than measured. The spectral shape of the vocal tract excitation strongly influences the observed envelope, and thus we can not guarantee that all vocal tract resonances will cause peaks in the observed spectral envelope, nor that all peaks in the spectral envelope are caused by vocal tract resonances.

In this thesis, the formant frequency estimation is based on modeling the speech signal as if it were generated by a particular kind of source and filter [43]. To find the best matching system, we use the Linear Prediction method. Linear prediction models the signal as if it were generated by a signal of minimum energy being passed through a purely-recursive Infinite Impulse Response (IIR) filter. Figure 3.9 shows the plot of the filter's frequency response by using linear prediction coding (LPC).

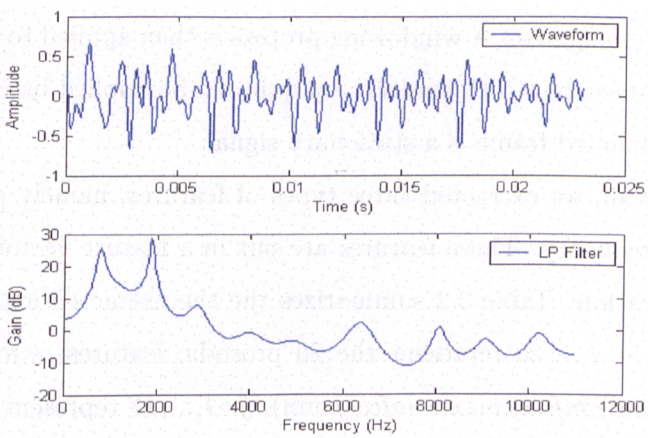


Figure 3.9: Formant frequency estimation. **Top:** waveform in time domain, **Bottom:** frequency response of the IIR filter using LPC

To find the formant frequencies from the filter, we need to find the locations of the resonances that make up the filter. This can be done by treating the filter coefficients as a polynomial and solving for the roots of the polynomial. In order to make the size of the features uniform, and achieve compromise between the imitation efficiency of the vocal tract system and the 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.

3.7 Summary

The performance of a human emotion recognition system is highly dependent on the accuracy and efficiency of the extracted features. As one of the major modalities of our proposed system, the accurate identification of speech features have critical influence on the recognition rate of the whole system. As we do not have prior knowledge of the significance of individual speech features, the efficiency of feature extraction can be better achieved by extracting a large number of different types of features and performing feature selection.

In this chapter, we present the methods that we used to extract audio features from the speech utterance to map the emotional speech onto the corresponding feature space. To reduce the negative effects of noise and useless silence, we perform preprocessing first. A windowing process is then applied to segment speech into short time frames, by which spectral analysis can be applied based on the assumption that the segmented frame is a stationary signal.

Overall, we extracted three types of features, namely prosodic, MFCC, and formant frequency. These features are put in a feature vector as the audio features for classification. Table 3.2 summarizes the the extracted audio features. In Table 3.2, $p(i)$, $i = 1,...,25$ represent the 25 prosodic features; $mfcc(j,\text{mean})$,$mfcc(j,\text{median})$, $mfcc(j,\text{std})$,$mfcc(j,\text{max})$, $mfcc(j,\text{min})$, $j=1,...,13$ represent the MFCC features; and $ff(k,\text{mean})$,$ff(k,\text{median})$, $ff(k,\text{std})$,$ff(k,\text{max})$, $ff(k,\text{min})$, $k=1,2,3$ represent the formant frequency features. The feature vector is organized in an order of prosodic,

Features	Dimension	Description
Prosodic	25	$[p(i),\ i = 1,...,25]$
MFCC	65	$[mfcc(j,\text{mean}), mfcc(j,\text{median}),\ mfcc(j,\text{std}),$ $mfcc(j,\text{max}), mfcc(j,\text{min}),\ j=1,...,13]$
Formant	15	$[ff(k,\text{mean}), ff(k,\text{median}),\ ff(k,\text{std}),$ $ff(k,\text{max}), ff(k,\text{min}),\ k=1,2,3]$
Overall Audio	105	$[p(1),...,p(25), mfcc(1,mean),...,$ $mfcc(13,min), ff(1,mean),..., ff(3,min)]$

Table 3.2: Summary table of audio feature extraction

MFCC, and formant frequency, with a total dimension of 105.

Chapter 4

Visual Feature Extraction

4.1 Introduction

FACIAL expression is another major factor in human emotion recognition. A face is one of the main output channels by which a person express emotional state. Different methods have been explored to perform facial expression analysis in the past, which can be roughly categorized into two groups. One is to treat the human face as a whole unit, and the other is to represent the face by prominent components, such as the mouth, eyes, nose, eyebrow, and chin. The analysis of components of the face will lose certain information, and thus we perform facial analysis by treating the face in a global scenario.

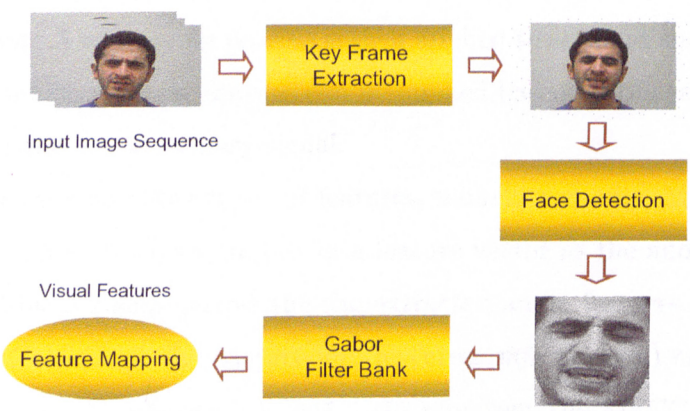


Figure 4.1: Flow of visual feature extraction

Figure 4.1 shows the visual feature extraction flow. In this thesis, to speed up the processing time, we use a key frame to represent the experimental subject's emotional state in a video clip. The key frame is the frame at which the corresponding speech has the highest amplitude. A face detection scheme based on HSV color model is then used to detect the face from the background. The facial expressions are represented by Gabor wavelet features.

4.2 Face detection

Different approaches of face detection have been studied in the past. Examples of these approaches include principal component analysis, skin color analysis, Hough transform, etc. Among these various methods, color analysis has been shown to be a promising prospect. One major advantage of color analysis is that the processing speed for color information is much faster than the other methods. This characteristic is very important for the purpose of real time human computer interaction.

In this thesis, we adopt a face detection scheme based on HSV color model. In a HSV color model, H and S components contain the chromatic information and V represents the luminance information. HSV color model corresponds closely to the human intuition on color. Figure 4.2 shows the model by the cone mode [45].

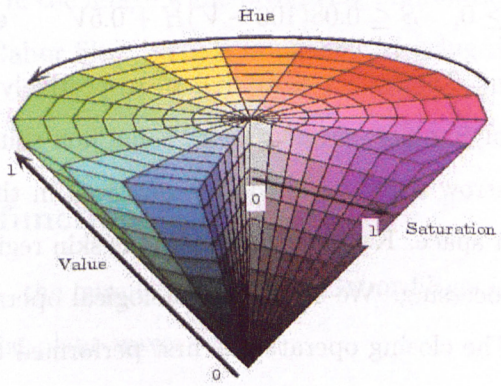


Figure 4.2: The HSV color space

Hue is the attribute of color and is measured by the angle around the vertical axis. It looks like a color wheel which is denoted by red, yellow, green, cyan, blue, magenta, and so on. Saturation is used to describe how pure a color is or how much white is added to a color. Value is a measure of the relative brightness.

The RGB components of an image can be converted to HSV color space by using the following formula:

$$H = \{H_1 \quad if B \leq G; \quad 360^0 - H_1 \quad if B > G\}, \tag{4.1}$$

where

$$H_1 = \cos^{-1}\{\frac{0.5[(R-G)+(R-B)]}{\sqrt{(R-G)(R-G)+(R-B)(G-B)}}\}, \tag{4.2}$$

$$S = \frac{max(R,G,B) - min(R,G,B)}{max(R,G,B)}, \tag{4.3}$$

$$V = \frac{max(R,G,B)}{255}. \tag{4.4}$$

We use the planar envelop approximation method [46] to approximate the human skin color. In planar envelope method, a pixel is considered as a skin pixel if the color of the pixel satisfies the following two conditions:

$$S \geq Th_s; \quad V \geq Th_v; \quad S \leq -H - 0.1V + 110; \quad H \leq -0.4V + 75, \tag{4.5}$$

$$\textbf{if} \quad H \geq 0, \quad S \leq 0.08(100 - V)H + 0.5V \qquad \textbf{else} \quad S \leq 0.5H + 35, \tag{4.6}$$

where Th_S and Th_V are set to 10 and 40 respectively [46].

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 falls into the skin color space. Keeping these spurious skin regions will yield negative effects to the latter processing. We apply morphological operation to implement the cleaning procedure. The closing operation is first performed to connect narrow gaps between skin regions, and then opening operation is applied to remove small isolated bulbs and separate the regions connected by thin strips. Finally, we perform the filling operation to remove the black isolated holes.

Figure 4.3 illustrates the procedure and operations of the applied face detection scheme. 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.

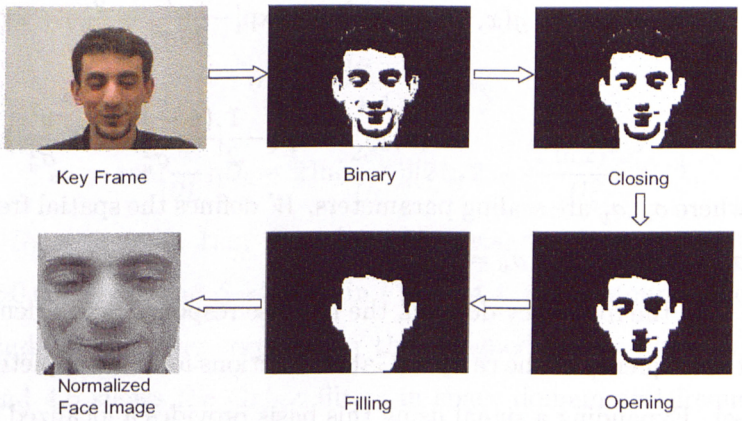


Figure 4.3: Procedure of the applied face detection scheme

4.3 Gabor wavelet representation

Using gabor wavelet features to represent facial expression has been explored and shown to be very efficient in the literatures [27]. It allows description of spatial frequency structure in the image while preserving information about spatial relations. In this thesis, the Gabor filter bank is implemented using the algorithm proposed in [47].

4.3.1 Gabor functions

In the space domain, the impulse response of Gabor filters is a Gaussian kernel modulated by a sinusoidal plane-wave:

$$g(x, y) = s(x, y)w(x, y), \tag{4.7}$$

where $s(x, y)$ is a complex sinusoidal known as the carrier, and $w(x, y)$ is a 2-D Gaussian-shaped function known as the envelope. A two dimensional Gabor function $g(x, y)$ and its Fourier transform $G(u, v)$ can be written as:

$$g(x,y) = (\frac{1}{2\pi\sigma_x\sigma_y}) \exp[-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}) + 2\pi jWx], \tag{4.8}$$

$$G(u,v) = \exp\{-\frac{1}{2}[\frac{(u-W)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}]\}, \tag{4.9}$$

where σ_x, σ_y are scaling parameters, W defines the spatial frequency of the sinusoidal, $\sigma_u = 1/2\pi\sigma_x$ and $\sigma_v = 1/2\pi\sigma_y$.

In the frequency domain, the impulse response is an oriented Gaussian centered at the frequency of the carrier. Gabor functions form a complete but nonorthogonal basis set. Expanding a signal using this basis provides a localized frequency description. A class of self-similar functions, referred to as Gabor wavelets, is considered. Let $g(x, y)$ be the mother Gabor wavelet, then this self-similar filter dictionary can be obtained by appropriate dilations and rotations of $g(x, y)$ through the generating function:

$$g_{mn}(x,y) = a^{-m} g(x', y'), \quad a > 1, \quad m, n = integer, \tag{4.10}$$

$$x' = a^{-m}(x\cos\theta + y\sin\theta), \tag{4.11}$$

$$y' = a^{-m}(-x\sin\theta + y\cos\theta), \tag{4.12}$$

where $\theta = n\pi/K$ and K is the total number of orientations. The scale factor a^{-m} is to ensure that the energy is independent of m.

4.3.2 Gabor filter dictionary design

The nonorthogonality of the Gabor wavelets implies that there is redundant information in the filtered images. Thus, the reduction of this redundancy has to be considered in the filter dictionary design. This can be solved by ensuring the half-peak magnitude support of the filter responses in the frequency spectrum to touch each other. If we let U_l and U_h denote the lower and upper center frequencies of

interest, K be the number of orientations, and S be the number of scales in the multiresolution decomposition, then the filter parameters σ_u, σ_v can be computed by the formulas below:

$$a = (\frac{U_h}{U_l})^{\frac{1}{S-1}}, \tag{4.13}$$

$$\sigma_u = \frac{(a-1)U_h}{(a+1)\sqrt{2\ln 2}}, \tag{4.14}$$

$$\sigma_v = \tan(\frac{\pi}{2k})[U_h - 2\ln(\frac{2\sigma_u^2}{U_h})][2\ln 2 - \frac{(2\ln 2)^2\sigma_u^2}{U_h^2}]^{-\frac{1}{2}}, \tag{4.15}$$

where $W = U_h$ and m=0, 1,..., S-1. In this thesis, the filter parameters used are U_h=0.4, U_l=0.05, K=6, and S=4 [47]. In Figure 4.4, the contours indicate the half-peak magnitude of the filter response in the designed Gabor filter dictionary [47]. Figure 4.5 and 4.6 shows the Gabor filters in space domain and frequency domain respectively, while an example of Gabor wavelet transformed face image is given in Figure 4.7.

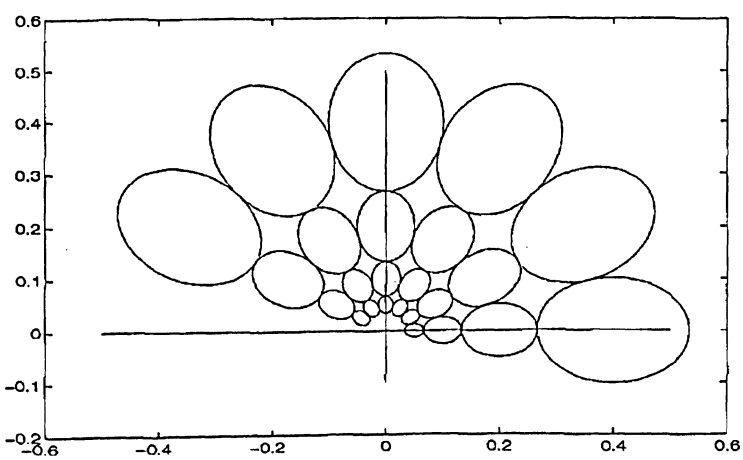


Figure 4.4: Contours of the designed Gabor filter dictionary

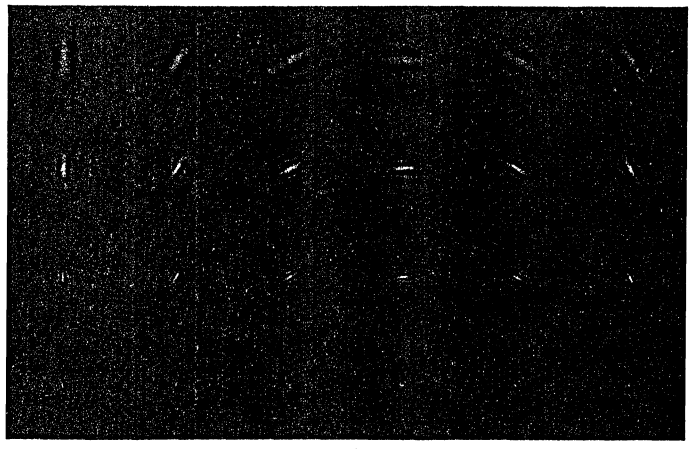


Figure 4.5: Gabor filters in space domain

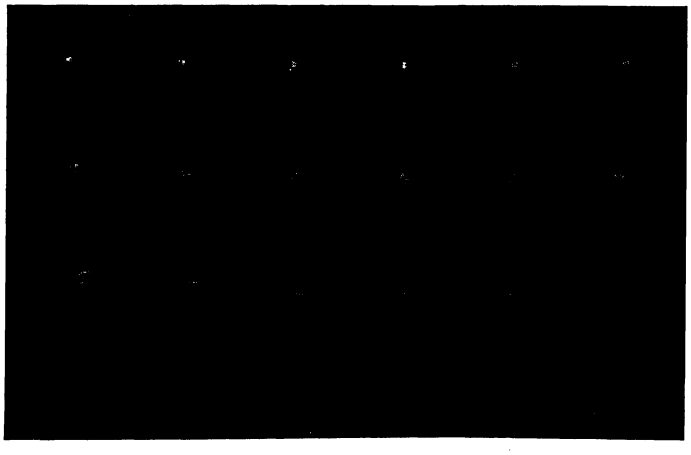


Figure 4.6: Gabor filters in frequency domain

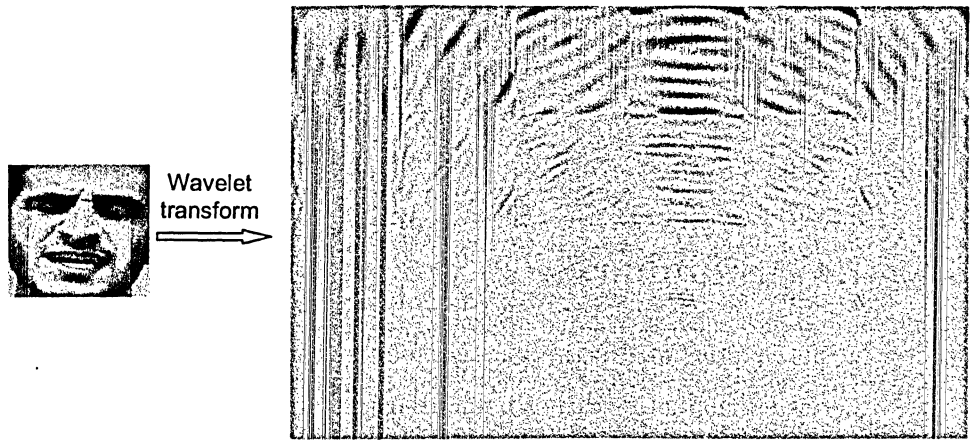


Figure 4.7: Example of Gabor wavelet transformed image

4.3.3 Feature representation

In section 4.2, we mentioned that the human face image is normalized to a grey-level image of size 128 × 128. If we take this image as the input image, and denote as $I(x,y)$, the Gabor wavelet transform of this image can be defined as:

$$W_{mn}(x,y) = \int I(x_1,y_1) g^*_{mn}(x-x_1, y-y_1) dx_1 dy_1, \tag{4.16}$$

where * indicates the complex conjugate. In this case, we obtain a very big coefficient matrix for each image. We use a total of 4 × 6=24 Gobor filters, and thus the size of the matrix is 128 × 128 × 24=393216. With a feature space of such a big size, the computation cost is very high, and thus it is not very suitable for a system which demands fast processing. We take the mean u_{mn} and standard deviation σ_{mn} of the magnitude of the transform coefficients of each filter to represent the features:

$$u_{mn} = \int\int |W_{mn}(x,y)| dx dy, \tag{4.17}$$

$$\sigma_{mn} = \sqrt{\int\int (|W_{mn}(x,y)| - u_{mn})^2 dx dy}. \tag{4.18}$$

A feature vector is constructed using u_{mn} and σ_{mn} as feature components. In this thesis, we use four scales S=4 and six orientations K=6, resulting in a feature vector of 48 dimension:

$$f_{Gabor} = [u_{00} \quad \sigma_{00} \quad u_{01} \quad \sigma_{01} \quad ..... \quad u_{35} \quad \sigma_{35}]. \tag{4.19}$$

4.4 Summary

In this chapter, we describe the methods that we used to extract features to represent the visual characteristics of human face at different emotional states. A key frame is first extracted from the image sequence. A face detection scheme based on HSV color model is employed to detect human face from the background. The detected face region is then normalized to a gray-level image of size 128 × 128. We utilize a Gabor filter dictionary design strategy to design a Gabor filter bank of 4 orientation and 6 scales. The normalized face image is passed through this filter bank, and the coefficients are computed. Finally, mean and standard deviation are computed from the output of each filter. These means and standard deviations are treated as the visual features. Totally, we extracted 48 visual features. These features in combination with the 105 audio features are put in a feature vector of size 153 for latter classification.

Chapter 5

Feature Selection and Classification

5.1 Introduction

THE recognition of human emotion is essentially a pattern recognition problem. To build a complete human emotion recognition system, a classifier needs to be applied to map the extracted features to the corresponding emotional state. This involves a machine learning procedure, by which the computer learns the characteristics of each emotional state. For a good classifier, the computer can successfully learn these characteristics and classify a new input correctly. However, which classifier we should select is a big challenge. In this thesis, we compared some popular classifiers, including Gaussian Mixture Model (GMM), K-nearest Neighbors (K-NN), Neural Network (NN), and Fisher's Linear Discriminant Analysis (FLDA).

We extracted a total number of 153 features from each sample to represent both the audio and visual information. With such a big dimensionality, the computational complexity is high. Moreover, not all the extracted features can contribute to the classification, some of the features might even cause negative effect due to possible dependency among the features. To reduce the dimensionality of the feature space, whilst maintaining the recognition accuracy, we need to apply feature selection algorithm. In this thesis, we also compared the performance of two feature selection algorithms, namely Principal Component Analysis (PCA) and a stepwise method.

5.2 Feature selection

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 the relevant subset from the original feature set is closely related to the "curse of dimensionality" problem in function approximation, where sample data points become increasingly sparse when 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 usually greater for implementing a high-dimensional mapping. To alleviate these problems, we reduce the dimensionality of input domain by choosing a relevant subset of features from the original set.

The ultimate goal of feature selection is to choose a number of features from the extracted feature set that yields minimum classification error. There exist several popular feature selection algorithms. Exhaustive search is an optimal algorithm which involves the search of all the possible combinations of the features. Although an optimal feature subset can be guaranteed, it is usually not a choice because of the high computational cost. The branch-and-bound feature selection algorithm [38] can be used to find the optimal subset of feature, but it requires the feature selection criterion function to be monotone, which means the value of the criterion function can never be decreased by adding new features to a feature subset. This is surely not the case when the sample size is relatively small. The wrapper method [48] is another choice of selecting the feature subset. However, it is very slow and computationally expensive because the whole data set needs to be analyzed repeatedly to evolve a feature set.

A good feature selection algorithm should have a tradeoff between computational complexity and accuracy. In [49], we investigate a combination of Sequential Forward Selection (SFS) with General Regression Neural Network (GRNN) for feature selection. But GRNN also has the disadvantage of high computational complexity [50]. In this section, we present the stepwise method that we adopted for feature

subset selection. A popular feature space reduction algorithm, Principal Component Analysis (PCA), is also presented for the purpose of comparative study.

5.2.1 Stepwise method

In this thesis, we perform feature selection using the stepwise method in SPSS (a trademark of SPSS Inc. USA). The stepwise method is a combination of forward and backward procedures. Forward procedure enters the variables in the list one by one until no more can be entered. The order of entering the variables is determined by the significance in the model. Backward procedure enters all the variables in the list in a single step, then removes the insignificant variables one by one until no more can be removed. However, both of these two procedures suffer a nesting problem. For instance, the selected 4-element subset is not contained in the selected 5-element subset. The combination of these two procedures can efficiently alleviate this nesting effect.

Stepwise method starts from one feature and adds one feature each time. The criterion to determine the inclusion or exclusion of a feature is the Mahalanobis distance. Mahalanobis distance is a measure of how much a case's value on the independent variables differ from the average of all cases. A large Mahalanobis distance identifies a case as having extreme values on one or more of the independent variables. Mahalonobis distance is defined as:

$$D_{mah} = (X_i - Y_i)' S^{-1} (X_i - Y_i), \tag{5.1}$$

where X_i and Y_i are the feature vectors, S^{-1} is the inverse covariance matrix. For each step, one feature is added to or removed from the selected feature subset to maximize the between-class Mahalanobis distance.

5.2.2 Principal component analysis

Principal component analysis is a popular transform-domain dimensionality reduction algorithm. The transformation is designed to represent the original features

by a reduced number of transformed features, whilst retaining most of the intrinsic information content of the original data [51].

To make the PCA work properly, the first step is to produce a data set whose mean is zero. This can be done by subtracting the mean from each of the data dimensions. The mean subtracted is the average across each dimension, and can be calculated as:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \tag{5.2}$$

where X denotes the entire set of samples in one dimension, n is the number of samples, X_i is the component of set X.

Then we need to compute the covariance matrix. Covariance is a measure to find out how much the dimensions vary from the mean with respect to each other. It can be calculated using the following equation:

$$cov(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})}{(n-1)}. \tag{5.3}$$

Covariance is always measured between 2 dimensions. for a m-dimension data set, $\binom{m}{2}$ different covariance values need to be calculated. To compute the covariance matrix, we need to calculate all the possible covariance values between all the different dimensions and put them in a matrix. The definition of the covariance matrix for a set of data with m dimensions is:

$$C^{m\times m} = [c_{i,j}], \quad c_{i,j} = cov(Dim_i, Dim_j), \tag{5.4}$$

where $C^{m\times m}$ is a matrix with m rows and m columns, and Dim_i is the ith dimension. For instance, the entry on row 2, column 3 of the covariance matrix is the covariance value calculated between the 2nd dimension and the 3rd dimension.

Since the covariance matrix is square, we can calculate the eigenvectors each with its associated eigenvalue for this matrix. These eigenvectors are orthogonal. The eigenvector with the highest eigenvalue is the first principal component of the data set. We can then order the eigenvectors by the associated eigenvalues, from the highest to the lowest. This gives the components in order of significance. The components with less significance can be ignored and thus dimension reduction can be achieved.

The final step is to form new data vectors by projecting the original data onto the principal component vectors. Once we have chosen the components that we wish to keep in our data and formed a feature vector by putting the eigenvectors in a matrix with the eigenvectors in the columns, we simply take the transpose of the vector and multiply it on the left of the original data set transposed.

$$X_{pca} = V_{eig} \times X_{original}, \tag{5.5}$$

where V_{eig} is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in rows, with the most significant eigenvector at the top. $X_{original}$ is the mean-adjusted data transposed. The new dimensionality-reduced data set X_{pca} can be used for latter classification.

5.3 Classification

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. In this thesis, we compared the performance of several popular classifiers, including Gaussian Mixture Model (GMM), K-nearest Neighbors (K-NN), Neural Network (NN), and Fisher's Linear Discriminant Analysis (FLDA).

5.3.1 Gaussian mixture model

GMM is widely used in speaker identification, verification, and speech recognition. GMM models the probability density function (*pdf*) of the data $P_{GMM}(y)$ as weighted sum of several different Gaussian density functions $p_k(y)$:

$$P_{GMM}(y) = \sum_{k=1}^{K} c_k p_k(y), \tag{5.6}$$

where c_k are the component probabilities, and K is the number of components. We use the Expectation Maximization (EM) algorithm to estimate the parameters of GMM [52]. The parameters including component probability c_k, mean μ_k, and covariance matrix $\sigma_{k,j}$ are estimated by using the following equations:

$$c_k^{i+1} = \frac{1}{N} \sum_{n=1}^{N} \psi_k(n), \tag{5.7}$$

$$\mu_k^{i+1} = \frac{\sum_{n=1}^{N} \psi_k(n) y_n}{\sum_{n=1}^{N} \psi_k(n)}, \tag{5.8}$$

$$\sigma_{k,j}^{(i+1)} = \frac{\sum_{n=1}^{N} \psi_k(n)(y_{n,j} - \mu_{k,j}^{(i)})^2}{\sum_{n=1}^{N} \psi_k(n)}, \tag{5.9}$$

where

$$\psi_k(n) = \frac{c_k^{(i)} p_k(y_n | \mu_k^{(i)}, \sum_k^{(i)})}{\sum_{j=1}^{K} c_j^{(i)} p_j(y_n | \mu_j^{(i)}, \sum_j^{(i)})}, \tag{5.10}$$

is a posteriori probability, N is the size of the feature vector, and i is the iteration number.

For classification, GMM is usually performed in a modular architecture, which involves a separate GMM being trained for each individual class. Figure 5.1 shows the architecture of the applied GMM classification scheme.

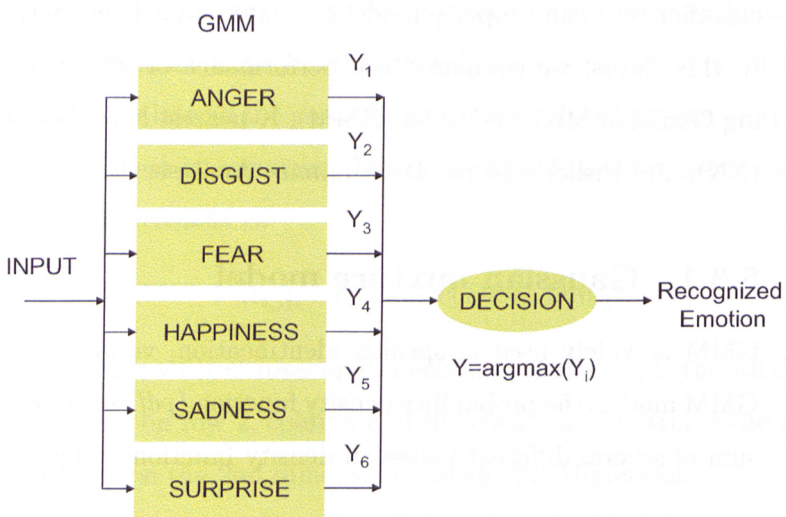


Figure 5.1: Architecture of the applied GMM scheme

An input signal is labeled the corresponding emotion with the maximum output:

$$Y = argmax(Y_i) \quad i = 1, 2, ..., 6. \tag{5.11}$$

5.3.2 K-nearest neighbors

K-nearest Neighbors is a non-parametric method for classification [53]. It assigns a class label to the unknown classification by examining its k nearest neighbors of known classification. Let $\mathbf{x} = \mathbf{x}_j, j = 1, 2, ..., n$ denotes the known n feature vectors, p denotes the number of different classes, $l = l_j, j = 1, 2, ..., n | l_j \in (1, 2, ..., p)$ represents the correspondent class labels, whilst the reference training samples can be written as:

$$S_n = \{(\mathbf{x}_1, l_1), (\mathbf{x}_2, l_2), ..., (\mathbf{x}_j, l_j), ..., (\mathbf{x}_n, l_n)\}. \tag{5.12}$$

For a new input vector $\mathbf{y}$, the k-nearest neighbors algorithm is to identify the subset of k feature vectors from S_n that lie closest to $\mathbf{y}$ with respect to a pattern similarity function $D(\mathbf{y},\mathbf{x})$. In this thesis, the employed pattern similarity function is Euclidean distance, which is defined as:

$$D(\mathbf{y},\mathbf{x}) = \sqrt{\sum_{i=1}^{n} |\mathbf{y}_i - \mathbf{x}_i|^2}. \tag{5.13}$$

If we use F_q to denote the frequency of each class label in the k nearest neighbors subset, the input vector can be classified by the following rule:

$$l_{out} = argmax\{F_q, \quad q = 1, 2, ..., p\}. \tag{5.14}$$

5.3.3 Neural network

Artificial neural networks [51] have seen an explosion of interest over the past few years, and are being applied across an extraordinary range of problem domains. Unlike the classical statistical classification 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 by using examples. Figure 5.2 shows the architecture of the employed neural network.

The back-propagation algorithm is used to perform the training. The number of neurons in the input layer is equal to the dimension of the feature vector. The number

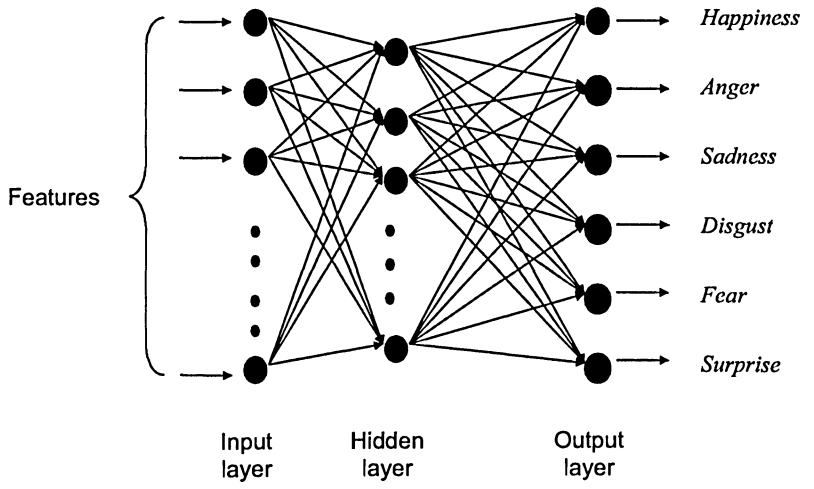


Figure 5.2: Architecture of the applied neural network

of neurons in the hidden hyperbolic tangent sigmoid layer N_{hidden} is determined by the following rule:

$$N_{hidden} = |\sqrt{N_{input} \times N_{output}}|, \tag{5.15}$$

where N_{input}, and N_{output} denote the number of neurons in the input layer and output layer respectively.

The hyperbolic tangent sigmoid transfer function is defined as:

$$\sigma(n) = (\frac{2}{1 + e^{-2n}}) - 1, \tag{5.16}$$

where $\sigma(n)$ is the output of a neuron in terms of the induced local field n.

The six neurons in the output linear layer corresponds to the six emotions. In the teaching process, we assign the value of 1 to the output neuron that corresponds to the same emotion as the input signal, and the value of 0 to all the other output layer neurons. In the simulating process, a new input is labeled the emotion with the greatest output.

5.3.4 Fisher's linear discriminant analysis

Linear discriminant analysis assumes the discriminant function $g(\mathbf{x})$ to be a linear function of data $\mathbf{x}$. In the case of c-class problem, the discriminant function is defined as [53]:

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}. \tag{5.17}$$

In Fisher's LDA, the weights $\mathbf{w}$ are estimated by maximizing Fisher's criterion function $J_F(\mathbf{w})$:

$$J_F(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}, \tag{5.18}$$

where S_b is the between-class scatter matrix, and S_w is the within-class scatter matrix.

If we use μ_i to represent the mean of class i, μ represents the mean of the overall data set, then S_b and S_w can be calculated as:

$$S_b = \sum_{i=1}^{C} \frac{n_i}{n} (\mu_i - \mu)(\mu_i - \mu)^T, \tag{5.19}$$

$$S_w = \sum_{i=1}^{C} \frac{n_i}{n} \sum_{x_k \in C_i} (x_k - \mu_i)(x_k - \mu_i), \tag{5.20}$$

where C is the number of classes, C_i denotes the samples in class i, n is the total number of samples, and n_i is the number of samples in class i. n_i/n represents a priori probability.

As our purpose is to find out $\mathbf{w}$, which will maximize the criterion $J_F(\mathbf{w})$, it can be deduced and found by solving the generalized eigenvalue problem:

$$J_F(\mathbf{w}) = \mathbf{w}^T S_b S_w^{-1} \mathbf{w}. \tag{5.21}$$

After the weights being calculated, we find a constant value w_{i0} for each discriminant function $g_i(\mathbf{x})$, which will maximize the separability between classes. For classification, as shown in Equation 5.17, the input data is classified into the class that gives the greatest discriminant function value.

5.4 Summary

In this chapter, we discuss the feature selection and classification algorithms that we used in this thesis. Table 5.1 summarizes the applied algorithms. The performance of the feature selection and classification algorithm is closely related to the inherent characteristics of the extracted features. By analyzing and comparing different approaches, it will help us to gain more insights into the investigated problem, and achieve higher recognition accuracy.

	Algorithms
Feature selection	Principal Component Analysis, Stepwise method
Classification	Gaussian Mixture Model, Neural Network, K-nearest Neighbors, Fisher's Linear Discriminant Analysis

Table 5.1: Feature selection and classification algorithms

Chapter 6

Emotion Recognition System

6.1 Introduction

SO far, we have discussed the feature extraction, feature selection and classification methods separately. An emotion recognition system is the combination of all these components. Figure 6.1 sketches an overview of the proposed recognition system. An input video sequence is passed through two different channels to extract the audio and visual characteristics of the sequence. These two streams of features are then fused into one stream, and a feature selection process is applied to find out significant features, whilst reducing the dimensionality, and thus the computational cost. 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 we performed were based on video samples from eight subjects (2 Italian, 2 Chinese, 2 Pakistani, 1 Persian, and 1 Canadian), speaking six different languages (English, Italian, Urdu, Punjabi, Mandarin, and Persian). 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. Table 6.1 details the experimental data on different emotions.

In this chapter, we first present the experimental results of applying different classification algorithms. We then compare different feature selection algorithms by

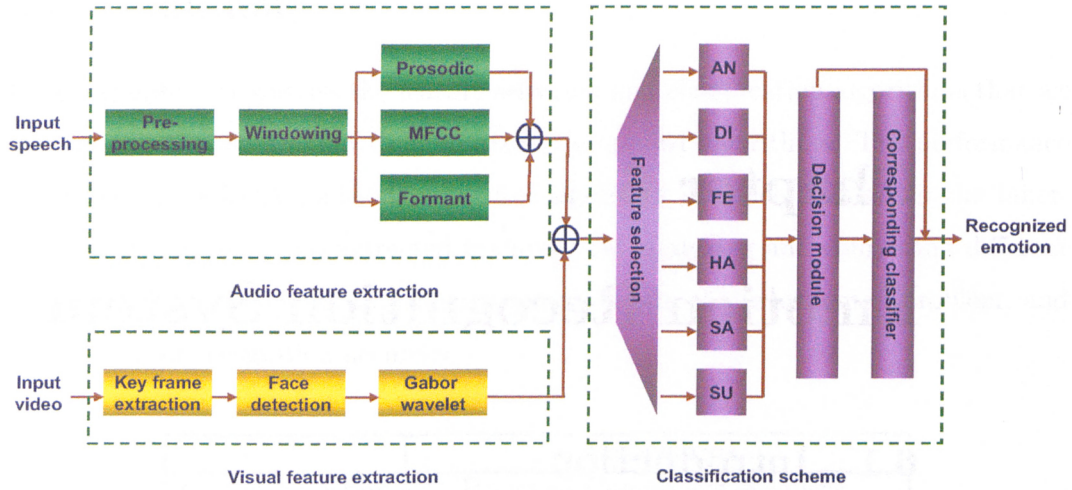


Figure 6.1: Overview of the emotion recognition system

Emotion	Training	Testing	Total
Anger	60	26	86
Disgust	59	21	80
Fear	56	22	78
Happiness	62	25	87
Sadness	59	21	80
Surprise	64	25	89
Total	360	140	500

Table 6.1: Experimental arrangement of the emotion database

using the best-performance classifier. A multi-classifier scheme based on the analysis of individual class and combinations of different classes is then introduced. Finally, leave-one-out cross validation results are also presented for comparative study.

6.2 Comparison of classification algorithms

In our first experiment, we performed classification on prosodic, audio, visual, and audiovisual features separately. We compared the performance of Gaussian Mixture Model (GMM), K-nearest Neighbors (K-NN), Neural Network (NN), and Fisher's Linear Discriminant Analysis (FLDA).

6.2.1 Experimental results

Gaussian mixture model

The GMM classifier is implemented in a modular architecture. A separate GMM is trained for each individual class. The parameters including the weights, mean, and standard deviation of each component are estimated by using EM algorithm. In out experiments, because we do not have a precise idea about the distribution of the data, we try a range of k value, so that the distribution of the data can be modeled as the sum of k Gaussian functions. By applying the GMM classification scheme on 25 Prosodic features, 105 audio features, 48 visual features, and 153 audiovisual features, we get the experimental results shown in Table 6.2, with the best recognition rate of each category shown in boldface.

k	Prosodic	Audio	Visual	Audiovisual
2	52.62%	58.57%	27.20%	63.27%
3	56.95%	62.11%	26.72%	72.37%
4	49.99%	**64.72%**	**28.28%**	58.56%
5	53.04%	57.51%	25.35%	62.24%
6	54.23%	61.65%	21.80%	60.69%
7	**60.60%**	61.66%	27.93%	60.23%
8	50.97%	56.97%	21.68%	**65.38%**
9	51.53%	57.86%	26.72%	55.24%
10	51.18%	55.49%	24.38%	57.91%

Table 6.2: Recognition results of GMM

K-nearest neighbors

In K-NN, an input signal is labeled the class that have the most samples in its k nearest neighbors. 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 our experiments, we do not set the k to a fixed value. We perform experiments on ten k values from one to ten, and the experimental results are shown in Table 6.3 with highest results in boldface.

k	Prosodic	Audio	Visual	Audiovisual
1	46.43%	46.43%	**35.71%**	57.14%
2	45.71%	50.71%	30.00%	50.71%
3	50.71%	55.71%	30.71%	57.14%
4	48.57%	55.71%	29.29%	57.14%
5	50.71%	58.57%	29.29%	57.14%
6	50.00%	55.71%	30.71%	60.00%
7	49.29%	61.43%	29.29%	**62.86%**
8	49.29%	**62.14%**	30.71%	57.86%
9	52.14%	61.43%	28.57%	59.29%
10	**55.00%**	59.29%	27.14%	60.00%

Table 6.3: Recognition results of K-NN

Neural networks

A three-layer feed-forward neural network is also investigated for classification. 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. Back-propagation algorithm is used to train the network. A new input is labeled the class that produces maximum output value. Table 6.4 shows the experimental results of applying neural network on prosodic, audio, visual and audiovisual features.

Classifier	Prosodic	Audio	Visual	Audiovisual
Neural Network	49.29%	51.43%	35%	56.43%

Table 6.4: Recognition results of neural network

Fisher's linear discriminant analysis

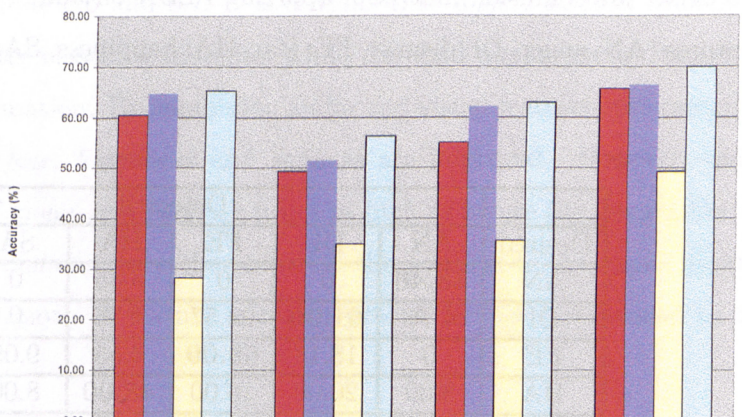
The applied FLDA classifier has six outputs corresponding to the six emotions, which is referred as a global classifier below. An input signal is labeled the class that gives the maximum output value. Table 6.5 shows the experimental results.

Classifier	Prosodic	Audio	Visual	Audiovisual
FLDA	65.71%	66.43%	49.29%	70.00%

Table 6.5: Recognition results of FLDA

6.2.2 Discussions

The comparison of the recognition results using different classification algorithms on prosodic, audio, visual, and audiovisual features is depicted in Figure 6.2.



	GMM	NN	K-NN	FLDA
Prosodic	60.60	49.29	55.00	65.71
Audio	64.72	51.43	62.14	66.43
Visual	28.28	35.00	35.71	49.29
Audiovisual	65.38	56.43	62.86	70.00

Figure 6.2: Recognition results of different classifiers

The experimental results show that the combination of audio and visual information performs better than either of them only. We compared the performance of GMM, K-NN, NN, and FLDA. It is obvious that FLDA outperforms the other classifiers. GMM and K-NN are statistical methods which are based on the estimation of probability density function. Neural networks estimate the optimal weight and bias values between different layers through a training process. They all need the training set to be large enough. In our case, the size of 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 reduce the performance.

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 the 105 dimension audio feature set, in which MFCCs and formant features are also included. However, this also demonstrate that phonetic features also contribute to the classification.

Another observation can be made is that the complementary relationship of audio and visual information enhance the performance of the system. Table 6.6, 6.7, and 6.8 detail the confusion matrix of applying FLDA on audio, visual, and audiovisual features (AN: anger, DI: disgust, FE: fear, HA: happiness, SA: sadness, SU: surprise).

	Detected					
Desired	AN	DI	FE	HA	SA	SU
AN	**88.46**	0	0	3.85	0	7.69
DI	4.76	**61.90**	28.57	4.76	0	0
FE	0	18.18	**59.09**	4.55	9.09	9.09
HA	8.00	20.00	16.00	**48.00**	8.00	0
SA	0	4.76	28.57	9.52	**57.14**	0
SU	16.00	0	4.00	0	0	**80.00**

Table 6.6: Confusion matrix of FLDA on audio features (all in %)

	Detected					
Desired	AN	DI	FE	HA	SA	SU
AN	**34.62**	0	42.31	7.69	15.38	0
DI	0	**47.62**	23.81	23.81	4.76	9.52
FE	18.18	9.09	**68.18**	0	0	4.55
HA	4.00	16.00	4.00	**52.00**	8.00	16.00
SA	9.52	4.76	14.29	14.29	**47.62**	9.52
SU	16.00	8.00	16.00	8.00	4.00	**48.00**

Table 6.7: Confusion matrix of FLDA on visual features (all in %)

	Detected					
Desired	AN	DI	FE	HA	SA	SU
AN	**76.92**	0	3.85	7.69	0	11.54
DI	0	**76.19**	19.05	4.76	0	0
FE	4.55	9.09	**77.27**	4.55	0	4.55
HA	4.00	16.00	8.00	**56.00**	16.00	4.00
SA	0	9.52	23.81	4.76	**61.90**	0
SU	12.00	0	12.00	4.00	0	**72.00**

Table 6.8: Confusion matrix of FLDA on audiovisual features (all in %)

From Table 6.6 and 6.7, it is obvious that *anger, disgust, sadness,* and *surprise* can be better classified in audio, while *fear* and *happiness* are better recognized by visual information. 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* is not as high as before the integration of data (Table 6.8). This demonstrates that some of the features might cause negative affects to the classification. A feature selection process is therefore needed to deal with this problem.

6.3 PCA *vs* Stepwise method

In our former experiments, we have shown that the integration of audio and visual information provides better recognition accuracy. However, with a feature space of 153 dimension, the computational cost is high. Moreover, not all the features can contribute to the classification. Some even cause negative effects. To reduce the dimensionality of the feature space, whilst achieving better recognition accuracy, we performed dimensionality reduction and feature selection. In this work, we compared two different algorithms, namely principal component analysis and a stepwise method.

6.3.1 Experimental results

Principal component analysis

In PCA, the dimensionality is reduced by eliminating less significant features with smaller eigenvalues in the transformed domain. In this study, we take the first m data from the newly generated data vectors, with m satisfy the following equation:

$$\frac{\sum_{i=1}^{m} E_i}{\sum_{j=1}^{N} E_j} \geq 90\%, \tag{6.1}$$

where E_x is the eigenvalue, and N is the dimension of the original feature space. Figure 6.3 depicts the curve of the eigenvalues produced in our experiments.

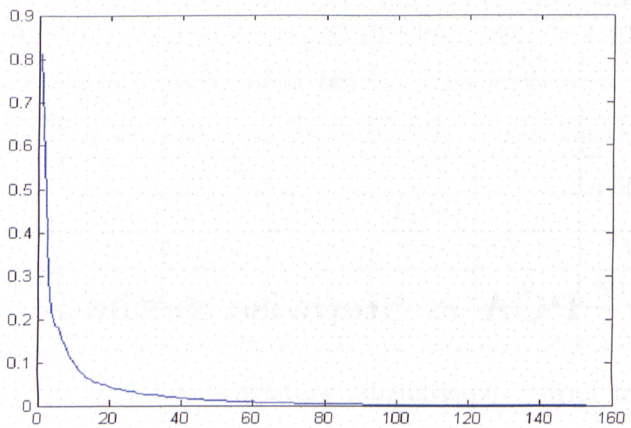


Figure 6.3: The eigenvalue curve produced in PCA

By using Equation 6.1, the sum of the eigenvalues of the first 40 data is greater than 90% of the sum of all the eigenvalues. It also can be observed from Figure 6.3 that the curve after 40 is almost flat and close to zero. Thus we take the first 40 data as the new features for classification. This method achieves 62.86% accuracy.

Stepwise method

We performed feature selection by using stepwise method. We selected 45 features from the original feature set, with the corresponding index numbers shown in Table

6.9. By using this method, the recognition rate was improved to 75.71%.

	Feature Index Number
Global FLDA Classifier	[2 3 7 19 21 22 24 25 26 27 28 34 43 48 55 68 78 80 91 107 108 109 111 113 114 116 118 119 121 122 124 127 128 130 131 132 133 134 135 137 139 143 146 148 149]

Table 6.9: Selected features using stepwise method

6.3.2 Discussions

Table 6.10 compares the dimensionality and recognition accuracy after applying feature selection algorithm. The recognition accuracy after applying PCA is actually lower than before feature space reduction. This is because some of the information is lost when the data is truncated. 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 from 153 to 45, and the recognition rate is also improved significantly.

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

Table 6.10: Comparison of feature selection algorithms

In comparison with the recognition rate of the original feature set, the selected feature subset achieves better overall accuracy, and specifically in *anger,happiness, sadness,* and *surprise* (Table 6.11). 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 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 emotion could have different significant features to separate from

all the other emotions or some specific emotions. This inspired us to perform feature selection and classification on individual emotions and combination of different emotions.

	Detected					
Desired	AN	DI	FE	HA	SA	SU
AN	**80.77**	0	3.85	7.69	0	7.69
DI	0	**61.90**	19.05	9.52	9.52	0
FE	0	18.18	**77.27**	4.55	0	0
HA	0	4.00	4.00	**80.00**	12.00	0
SA	0	4.76	14.29	4.76	**76.19**	0
SU	16.00	0	4.00	4.00	0	**76.00**

Table 6.11: Confusion matrix of FLDA on stepwise method selected audiovisual features (all in %)

6.4 The multi-classifier scheme

In this thesis, an adaptive multi-classifier scheme involving the analysis of individual class and combinations of different classes is proposed. As FLDA is shown to be performing better than the other classifiers, the individual classifiers in this multi-classifier scheme are based on FLDA. Figure 6.4 shows the architecture of the adaptive multi-classifier scheme.

6.4.1 Experimental results

We built six one-against-all (OAA) classifiers first, which is represented as " AN, DI, FE, HA, SA, SU" separately in Figure 6.4. In the training process, 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". A value greater than 50% represents the sample is classifier as "anger", while less than 50% represents "non-anger".

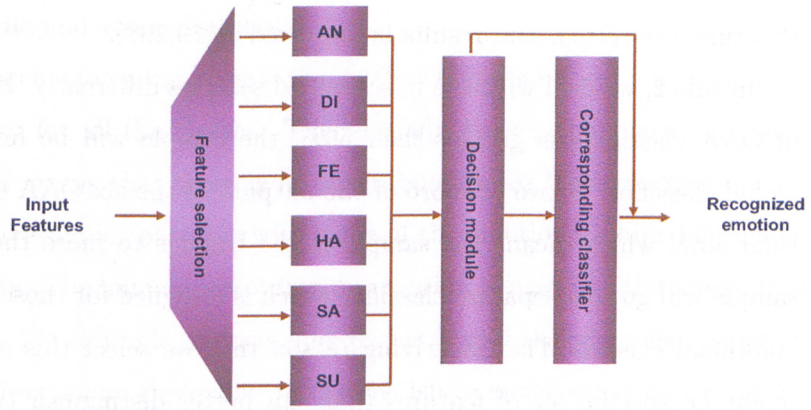


Figure 6.4: Architecture of the multi-classifier scheme

Feature selection was performed to find significant features for individual class. The applied feature selection algorithm is the stepwise method, which has been shown to perform better. Table 6.12 summaries the selected features for individual emotion.

	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 67 108 113 118 127 132 133 137]
HA	[1 3 6 32 49 50 55 56 65 74 75 78 80 98 99 108 109 111 116 122 124 128 130 131 132 133 135 138 142 145 146 148 151]
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]

Table 6.12: Selected features for individual class

We take the output of each OAA classifier 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 exceeds 50%, or two or more

greater than 50%, will go to the global classifier for further classification. By using this rule, the recognition results 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%, 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 that we select this scheme is that, there might be specific set of features that can better distinguish two or more different emotions. If a sample is misclassified as two or more emotions, this specific classifier that dedicate to the corresponding emotions will help to distinct minor difference in these emotions. 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 all the experiments, stepwise method is used to select the best feature subset. This system achieves 82.14% accuracy, with the confusion matrix shown in Table 6.13.

	Detected					
Desired	AN	DI	FE	HA	SA	SU
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 6.13: Confusion matrix of the multi-classifier scheme (all in %)

6.4.2 Discussions

From Table 6.12, we can find that the significant features to distinct one individual emotion from the other emotions are different. Some of the features selected in a global scenario are redundant (Table 6.9), 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.

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, by which means that there is no sharp boundaries between emotions. One emotion might have similar patterns with some of the emotions, while different patterns with the others. 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 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).

Taking advantage of the individual class based analysis, the recognition rate improves significantly. In our proposed multi-classifier scheme, rule 2 achieves higher accuracy than rule 1. This demonstrates that the significant features to distinguish 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.

6.5 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 as the training data. This procedure continues until all the individual samples in the entire data set has 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%, with the confusion matrix shown in Table 6.14.

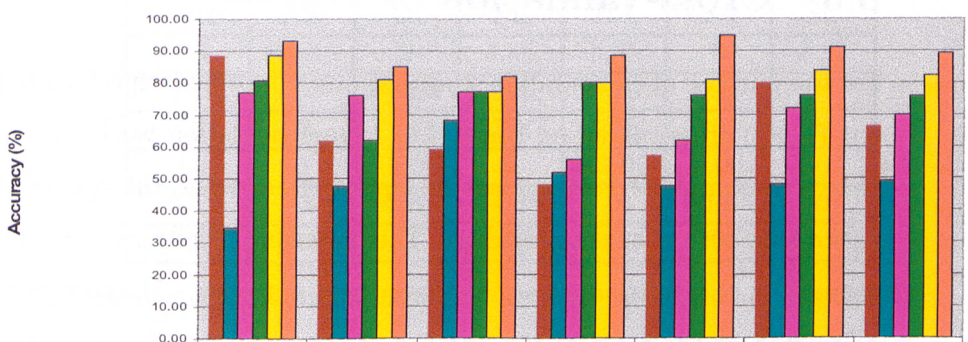
	Detected					
Desired	AN	DI	FE	HA	SA	SU
AN	**93.00**	1.20	0	1.20	0	4.70
DI	0	**85.00**	7.50	6.30	1.30	0
FE	2.60	7.70	**82.10**	2.60	3.80	1.30
HA	0	4.60	3.40	**88.50**	2.30	1.10
SA	0	3.80	1.30	0	**95.00**	0
SU	2.20	0	3.40	3.40	0	**91.00**

Table 6.14: Confusion matrix of LOO cross-validation (all in %)

6.6 Summary

In this chapter, we present the proposed emotion recognition system. Extensive experiments were conducted to test the effectiveness of our system. We compared the performance of different classifiers first. Based on the experiments, we select Fisher's Linear Discriminant Analysis (FLDA) as the classifier. Feature selection and classification schemes are then applied to enhance the system.

The recognition results using different features and classification schemes based on FLDA are summarized in Figure 6.5.



	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Overall
Audio	88.46	61.90	59.09	48.00	57.14	80.00	66.43
Visual	34.62	47.62	68.18	52.00	47.62	48.00	49.29
Audiovisual	76.92	76.19	77.27	56.00	61.90	72.00	70.00
Stepwise	80.77	61.90	77.27	80.00	76.19	76.00	75.71
Multi-classifier	88.46	80.95	77.27	80.00	80.95	84.00	82.14
LOO	93.00	85.00	82.10	88.50	95.00	91.00	89.20

Figure 6.5: Comparison of recognition results

The results demonstrate that the combination of audio and visual information achieves better overall accuracy than either of them only. The applied stepwise method efficiently reduce the dimensionality of the feature space, whilst achieving better recognition accuracy. The proposed adaptive multi-classifier scheme produces noticeable improvement in both overall and individual class recognition accuracy.

By using leave-one-out cross-validation method, we achieve very promising results. 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.

Chapter 7

Conclusions

7.1 Summary of thesis

IN this thesis, we present an emotion recognition system to recognize human affective states from audiovisual information. The proposed system was tested over a language, speaker, and context independent emotional video database. Our objective is to investigate the universal nature of emotion and its vocal and facial expressions, and thus enhance the human computer intelligent interaction.

In audio feature extraction, in order to identify the features that can truly represent the vocal characteristics of human speech for different emotions, we examine both prosodic and phonetic features. In visual feature extraction, we extract a key frame from the image sequence first, a HSV color model based face detection scheme is then applied to segment the face region from the background. We perform multiresolution Gabor wavelet analysis on the face image for the computational mapping between emotions and facial expressions. The motivation of adopting this visual feature extraction method is to facilitate the processing speed of the system, and thus for potential real life applications.

To find a classifier that can better model the data and map the features to the corresponding emotions, we compared different classifiers, including parametric classifier Gaussian Mixture Model (GMM), non-parametric classifier K-nearest Neighbors (K-NN), non-linear classifier Neural Network (NN), and linear classifier Fisher's Linear Discriminant Analysis (FLDA). FLDA is selected as classifier through experiments.

Our proposed recognition system appears to be a linear problem. However, in specific emotional class, the performance of audiovisual features is worse than audio features only. This demonstrates that some of the features might cause negative effects. In order to identify the significant features, whilst reducing the dimensionality of the feature space, we perform feature selection. The criteria for the selection of a feature selection algorithm is to get a tradeoff between the efficiency and computational complexity. We compared the performance of two feature selection algorithms, principal components analysis and a stepwise method. Experimental results show that the stepwise method is capable of identifying a highly relevant subset of features, reducing the dimensionality of the feature space, and improving the recognition accuracy.

In this work, we also propose a multi-classifier scheme to further analyze and enhance the classification accuracy. This multi-classifier scheme takes advantage of the analysis of significant features in individual class and that to distinguish any combinations of any classes. Experiments demonstrate the effectiveness of the proposed scheme.

We perform extensive experiments to compare and evaluate the adopted and proposed approach. Based on the implementation and the results, some major conclusions can be drawn as follow:

- Phonetic features also contribute to the classification of human emotion. Although prosodic features are a very powerful indicator of human emotional states, the MFCCs and formant frequency features are extracted to mimic the human hearing system and vocal tract, and thus human perception of emotional speech can be better simulated. In our experiments, we perform feature selection to find out significant features. In all the cases, some MFCCs and formant features are selected as significant features for classification.

- The combination of audio and visual information performs better than either of them only. The complementary relationship of these two modalities on different emotions help to achieve higher recognition accuracy. For instance, in a facial expression analysis system, the performance might be greatly affected by

facial hair, skin color etc. However, this can be compensated from the audio information.

- Although language and cultural background might have some influence on the way in which people express their emotions, our proposed system shows promising results, and demonstrated that the emotional expressions can be identified beyond these boundaries.

- The recognition accuracy can be greatly improved by partitioning the classification problem into finer analysis, which is based on individual class and combinations of specific emotions. This is due to the fact that the features selected to distinguish a set of emotions might be not distinctive for another set of emotions. Our experiments show that the features selected for individual class is different from that of being selected in a global scenario.

7.2 Contributions

In this thesis work, we investigate the advantage and effectiveness of combining multimodality information in human emotion recognition. It opens up the potential and possibility of analyzing other human computer interaction (HCI) problems. HCI is a multi-channel behavior. The combination of different modalities conveys more information about the interface, and more importantly, the complementary relationship between different modalities will help to enhance the performance of the system.

Rather than performing analysis on all the classes, we also analyze the significant features on an individual class basis. It helps us to obtain more detailed insight into individual emotion and the way to separate specific emotions. We perform experiments in a logical and systematic manner. Our proposed system achieves very promising results by using the proposed classification scheme. Considering a more generic application, our system was tested using a very versatile database, in which samples from different subjects, speaking different languages were collected. We have considered the processing speed in audio and image processing, and we also performed

feature selection to reduce the dimensionality, and thus the computational cost. All these demonstrates that a real life application can be expected.

7.3 Future research

In our research, we studied the vocal and facial expressions and propose a system to recognize these expressions. In this section, we outline several proposals which may further enhance the performance of the system:

- We perform emotion analysis on six basic emotions. However, human emotional states do not have a sharp boundary. Some of the emotions are a combination of different emotions. For instance, human can express different surprise, sometimes combined with happy, and sometimes with fear. For a natural human computer interface, the computer needs to recognize and analyze these situations. One proposal is to categorize emotion into a wide range of classes. Another might be giving different weights to the basic emotional elements, and human can understand the potential components of an expression.

- From our experiments, the visual feature based classification accuracy is low. This demonstrates that the visual feature representation is not strong enough. In the future , we are going to work on a hybrid approach which incorporates the analysis of the whole face and the prominent parts of the face, as well as dynamic features in video sequence

- The emotion analysis are performed on a video signal that have only one emotion. In real life applications, the user's emotion is changing frequently. A scheme needs to be proposed to separate the varying of human emotion in one continues interaction. Furthermore, the system should capable of detecting the presence of the user automatically.

- We fuse the audio and visual data in a simple manner. In the future, we are going to devise a fusion algorithm which will better utilize the complementary relationship of the two modalities.

- We use one classifier for all the six emotions. However, different class might have different classification algorithm that can better model the data. An investigation based on these scenarios will help to improve the efficiency of the system.

Bibliography

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

[2] C. Calhoun, R. C. Solomon, What is an Emotion, New York: Oxford University Press, 1984

[3] C. Darwin, The Expression of Emotions in Man and Animals, John Murray, 1872, reprinted by University of Chicago Press, 1965

[4] *www.driesen.com/limbicsystempic21.jpg*

[5] P. Ekman, M. O'Sullivan, "The role of context in interpreting facial expression: Comment on Russell and Fehr", Journal of Experimental Psychology, General, vol. 117, pp. 86-88, 1987

[6] A. J. Calder, A. D. Lawrence, and A. W. Young, "Neuropsychology of fear and loathing", Nature Reviews Neuroscience, vol. 2, pp. 352-363, 2001

[7] R. R. Cornelius, The Science of Emotion. Research and Tradition in the Psychology of Emotion. Upper Saddler River: Prentice Hall, 1996

[8] P. Ekman, "Strong evidence for universals in facial expressions: A reply to Russelll's mistaken critique", Psychological Bulletin, vol. 115, pp. 268-287, 1994

[9] L. C. De Silva, T. Miyasato and R. Nakatsu, "Facial emotion recognition using multi-modal information", Proceedings of IEEE International Conference on Information, Communications and Signal Processing, vol. 1, pp. 397-401, Singapore, September 1997

[10] R. Cowie, E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech", Proceedings of 4th International Conferences on Spoken Language Processing, vol. 3, pp. 1989-1992, Philadelphia, USA, October 1996

[11] N. Amir, S. Ron, "Toward an automatic classification of emotions in speech", Proceedings of 5th International Conference on Spoken Language Processing, vol. 3, pp. 555-558, Sydney, Australia, November/December 1998

[12] J. Sato, S. Morishima, "Emotion modeling in speech production using emotion space", Proceedings of IEEE International Workshop on Robot and Human Communication, pp. 472-477, Tsukaba, Japan, November 1996

[13] F. Dellaert, T. Polzin and A. Waibel, "Recognizing emotion in speech", Proceedings of the 4th International Conferences on Spoken Language Processing, vol. 3, pp. 1970-1973, Philadelphia, USA, October 1996

[14] J. Nicholson, K. Takabashi and R. Nakatsu, "Emotion recognition in speech using neural networks", Neural Computing and Applications, vol. 9, pp. 290-296, 2000

[15] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Classifying emotions in human-machine spoken dialogs", Proceedings of International Conference on Multimedia and Expo, vol. 1, pp. 737-740, Switzerland, August 2002

[16] T. L. Nwe, F. S. Wei and L. C. De Silva, "Speech based emotion classification", Proceedings of IEEE Region 10 Conference on Electrical and Electronics Technology, vol. 1, pp. 297-301, Singapore, August 2001

[17] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speech signals", Proceedings of European Conference on Speech Communication and Technology, pp. 125-128, Geneva, Switzerland, September 2003

[18] B. Schuller, G. Rand M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief net-

work architecture", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 577-580, Montreal, Canada, May 2004

[19] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 593-596, Montreal, Canada, May 2004

[20] P. Ekman and W. V. Friesen, Facial Action Coding System (FACS): Manual, Palo Alto: Consulting Psychologists Press, 1978

[21] G. Donato, M. S. Barlett, J. C¿ Hager, P. Ekman and T. J. Sejnowski, "Classifying facial actions", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, pp. 974-989, October 1999

[22] D. Yang, T. Kunihiro, H. Shimoda, and H. Yoshikawa, "A study of real-time image processing method for treating human emotion by facial expression", Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, vol. 2, pp. 360-364, Japan, October 1999

[23] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis", Psychophysiology, vol. 36, pp. 253-263, 1999

[24] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey", Pattern Recognition, vol. 25, pp. 65-77, 1992

[25] M. Pantic and L. J. M. rothkrantz, "Automatic analysis of facial expressions: The state of the art", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 1424-1445, 2000

[26] K. Haneda, T. Muraguchi and O. Nakamura, "The recognition of faical expressions using expert system", Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering, vol. 2, pp. 1195-1198, Quebec, Canada, May 2003

[27] 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, pp. 202-207, France, March 2000

[28] M. Pantic and L. J. M. rothkrantz, "Facial action recognition for facial expression analysis from static face image", IEEE Transactions on Systems, Man, and Cybernetics-Part B:Cybernetics, vol. 34, pp. 1449-1461, 2004

[29] L. C De Silva and S. C. Hui, "Real-time facial feature extraction and emotion recognition", Proceedings of 4th International Conference on Information, Communications and Signal Processing, vol. 3, pp. 1310-1314, Singapore, December 2003

[30] I. Cohen, N. Sebe, Y. Sun, M. S. Lew and T. S. Huang, "Evaluation of expression recognition techniques", Proceedings of International Conference on Image and Video Retrieval, pp. 184-195, Urbana, IL, USA, July 2003

[31] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks", IEEE Transactions on Systems, Man and Cybernetics-Part B:Cybernetics, vol. 34, pp. 1588-1595, 2004

[32] D. Kim and Z. Bien, "Fuzzy neural networks(FNN)-based approach for personalized facial expression recognition with novel feature selection method", Proceedings of IEEE International Conference on Fuzzy Systems, vol. 2, pp. 908-913, St. Louis, MO, USA, May 2003

[33] M. Song, C. Chen, and M. You, "Audio-visual based emotion recognition using tripled hidden Markov model", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 877-880, Montreal, Canada, May 2004

[34] L. C. De Silva, P. C. Ng, "Bimodal emotion recognition", Proceedings of 4th

IEEE International Conference on Automatic Face and Gesture Recognition, pp. 332 - 335, France, March 2000

[35] L. S. Chen, H. Tao, T. S. Huang, T. Miyasato, and R. Nakatsu, "Emotion recognition from audiovisual information", Proceedings of IEEE 2nd Workshop on Multimedia Signal Processing, pp. 83-88, California, USA, December 1998

[36] H. Go, K. Kwak, D. Lee, and M. Chun, "Emotion recognition from the facial image and speech signal", Proceedings of SICE 2003 Annual Conference, vol. 3, pp. 2890-2895, Japan, August 2003

[37] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 34-58, January 2002

[38] P. M. Narendra and K. fukunaga, "A branch and bound algorithm for feature subset selection", IEEE Transactions on Computers, vol. 26, pp. 917-922, September 1977

[39] J. Kittler, "Feature set algorithms", Pattern Recognition and Signal Processing, pp. 41-60. Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, 1978

[40] G. Bailly, C. Benoit, and T. R. Sawallis, Talking Machines: Theories, Models, and Designs. Elsevier Science Publishers, Amsterdam: 1992

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

[42] C. Becchetti, and L. P. Ricotti, Speech Recognition: Theory and C++ Implementation. Toronto: John Wiley and Sons, 1999

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

[44] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking", Proceedings of the Symposium on Time Series Analysis, pp. 209-243, June 1962

[45] *www.aurora – soft.co.uk/images/hsv.jpg*

[46] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis", IEEE Transactions on Multimedia, vol. 1, pp 264-277, September 1999

[47] 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, pp. 837-842, August 1996

[48] R. Kohavi and G. H. John, "Wrappers for feature subset selection", Artificial Intelligence, vol. 97, pp. 273-324, 1997

[49] W. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech", Proceedings of IEEE International Symposium on Circuits and Systems, vol. 2, pp. 181-184, Vancouver, May 2004

[50] D. F. Specht, "A general regression neural network", IEEE Transactions on Neural Networks, vol. 2, pp. 568-576, 1991

[51] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1999

[52] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models", Technical Report ICSI-TR-97-021, International Computer Science Institute, Berkeley, CA, April 1998

[53] S. Theodorids, and K. Koutroumbas, Pattern Recognition (2nd Edition), Elsevier Science (USA), 2003

Appendix A

List of Publications

The publications based on the work of this thesis are listed below:

Journal Papers

- Waqas Bhatti, Yongjin Wang, and Ling Guan, "Human emotion recognition in speech using artificial neural networks", submitted to IEEE Transactions on Neural Networks. December, 2003
- Yongjin Wang, and Ling Guan, "Language, speaker, and context independent recognition of human emotion from speech signals", submitted to IEEE Transactions on Circuits and Systems II. October, 2004

Conference Papers

- Waqas Bhatti, Yongjin Wang, and Ling Guan, "A neural network approach for human emotion recognition in speech", Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), Page(s): II- 181-4 Vol.2, Vancouver, May 2004
- Yongjin Wang, and Ling Guan, "An investigation of speech-based human emotion recognition", Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP), Page(s): 15-18, Siena, Italy, September 29- October 1, 2004

- Yongjin Wang, and Ling Guan, "Recognizing human emotion from audiovisual information", accepted by IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, USA, March 2005