Summer 8-2014

Multi-Sensory Emotion Recognition with Speech and Facial Expression

Qingmei Yao
University of Southern Mississippi

The University of Southern Mississippi

MULTI-SENSORY EMOTION RECOGNITION

WITH SPEECH AND FACIAL EXPRESSION

by

Qingmei Yao

Abstract of Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

August 2014

ABSTRACT

MULTI-SENSORY EMOTION RECOGNITION

WITH SPEECH AND FACIAL EXPRESSION

by Qingmei Yao

August 2014

Emotion plays an important role in human beings' daily lives. Understanding emotions and recognizing how to react to others' feelings are fundamental to engaging in successful social interactions. Currently, emotion recognition is not only significant in human beings' daily lives, but also a hot topic in academic research, as new techniques such as emotion recognition from speech context inspires us as to how emotions are related to the content we are uttering.

The demand and importance of emotion recognition have highly increased in many applications in recent years, such as video games, human computer interaction, cognitive computing, and affective computing. Emotion recognition can be done from many sources including text, speech, hand, and body gesture as well as facial expression. Presently, most of the emotion recognition methods only use one of these sources. The emotion of human beings changes every second and using a single way to process the emotion recognition may not reflect the emotion correctly. This research is motivated by the desire to understand and evaluate human beings' emotion from multiple ways such as speech and facial expressions.

In this dissertation, multi-sensory emotion recognition has been exploited. The proposed framework can recognize emotion from speech, facial expression, and both of them. There are three important parts in the design of the system: the facial emotion

recognizer, the speech emotion recognizer, and the information fusion. The information fusion part uses the results from the speech emotion recognition and facial emotion recognition. Then, a novel weighted method is used to integrate the results, and a final decision of the emotion is given after the fusion.

The experiments show that with the weighted fusion methods, the accuracy can be improved to an average of 3.66% compared to fusion without adding weight. The improvement of the recognition rate can reach 18.27% and 5.66% compared to the speech emotion recognition and facial expression recognition, respectively. By improving the emotion recognition accuracy, the proposed multi-sensory emotion recognition system can help to improve the naturalness of human computer interaction.

The University of Southern Mississippi

MULTI-SENSORY EMOTION RECOGNITION

WITH SPEECH AND FACIAL EXPRESSION

by

Qingmei Yao

A Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved:

Dr. Chaoyang Zhang
Committee Chair

Dr. Shaoen Wu

Dr. Nan Wang

Dr. Zhaoxian Zhou

Dr. Huiqing Zhu

Dr. Maureen Ryan
Dean of the Graduate School

August 2014

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. Chaoyang Zhang, for his kindness, help, and advisement during the last year of my Ph.D. study. I would also like to give my earnest appreciation to my advisor, Dr. Shaoen Wu, who has given me the opportunity to study and research at The University of Southern Mississippi from spring 2010 to 2013. Thanks for his support, patience, encouragement, and great research guidance during my Ph.D. program. In addition, I am grateful to other committee members of my dissertation Drs. Nan Wang, Zhaoxian Zhou, and Huiqing Zhu for their valuable remarks and suggestions on my dissertation. Finally, I would like to thank all my family members, my parents, my sister, my husband – Bin and my lovely daughter – Iris. Without their selfless support and encouragement, I could not complete the research and dissertation.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF ILLUSTRATIONS

Figure

CHAPTER I

INTRODUCTION

Currently computers are playing a more and more important role in many fields of human beings' lives. Traditional interfaces of computers like the keyboard and the mouse are not able to satisfy human beings' requirements. So, offering a natural interaction between computer and human is highly needed.

So far, human computer interaction is far less natural than the interaction among human beings. It is not possible yet to do face-to-face communication through human computer interaction completely. To improve the naturality, computers should be able to support the way human beings communicate with each other. The essential requirement is that they should have the ability to acquire and show emotions. They should also have the ability to recognize and understand the emotions of the human counterpart in order to have a more natural interaction like the interaction between human beings.

Emotion is one of the most complicated, physical, and psychological behavior of human according to the theories of emotion of Meyers (2004). In human daily social life, knowing the emotional feeling of the counterpart is critical and intuitive. However, when it comes to the computer, it is much harder to determine the emotion of its counterpart.

In human communication, there are many aspects such as speaking, body gestures, eye contact, and facial expressions. The basic units of verbal communication are words in speech and the basic units of the nonverbal communication include the facial expression, body movements, and gestures as well as the eye contact. Sometimes, speech itself is sufficient for communication, take the phone call for example. Nonverbal communication units are also important, especially in face-to-face communication. Moreover, the

emotions reflected from facial expression are far more intense than words. So, to approach more natural human computer interaction, the computer should have the ability to recognize the emotion from speech, facial expression, or body gestures. In this dissertation, we proposed a multi-sensory emotion recognition framework to recognize emotion from speech and facial expression. Experiments show that multi-sensory emotion recognition can help the computer understand emotions of human beings more accurately. Hence, the proposed multi-sensory emotion recognition can help to improve the human computer interaction more natural.

Emotions and Emotion Recognition

What is emotion? From the psychological viewpoint, emotion is often determined as a complex state of feeling which represents the physical and psychological changes that can affect human beings' thought and behavior. Emotions can be reflected from voices, speeches, hand and body gestures, as well as facial expressions that play an important role in our daily life. When a person smiles, it may indicate that the person is happy at that moment, while if one is frowning it may establish that the emotion of this person is sadness or anger.

The number of emotions that human beings use in everyday life is hard to specify because of different situations and environments. It is widely accepted by psychological theory that the emotions of humans can be classified into six general categories happiness, anger, sadness, fear, disgust, and surprise. However, at different time human expresses different moods about how they are feeling by showing different facial actions and gestures during communications. The emotional expressions used are much more, and some of them are combinations of more than one.

With the rapid development of the computer and the Internet, understanding emotions and recognizing how to react to people's expressions are becoming important in video games, human computer interaction, cognitive computing, and affective computing. Thus, emotion recognition and its application gain more and more popularity in both scientific research field and industry field. Nevertheless, to observe or interpret human emotion is not an easy work for computers since emotion itself is complicated.

Currently, emotion recognition can be done via many ways including facial expression, voice, body gestures, as well as body movements, in which facial emotion recognition and speech emotion recognition get increasing interest in designing human machine interaction. Oral communication is a rich source of information for emotion recognition when people communicate with others. Moreover, sometimes it is not important what we have spoken, but how we spoke. The facial expression is the most visible form of emotional communication, but it is also most easily controlled by the speakers in different social environments when compared to speech and other kinds of expression. Thus, emotion recognition by a single way, such as facial expression or speech, may not reflect the person's emotional state correctly.

To overcome the limitation of a single way emotion recognition, multimodal emotion recognition was proposed theoretically. Schuller, Lang, and Rigoll (2002) discussed multimodal emotion recognition by analyzing the user's speech signal and haptical interaction on a touch screen or via mouse. In their work not only the common prosodic speech features (pitch and energy), but also the semantic and intention based features (wording, degree of verbosity, temporal intention and word rate, as well as the history of user utterances) were used. Multimodal emotion recognition from facial

expressions, body gestures, and speech was proposed by Caridakis et al. (2007) in which the Bayesian Classifier was used to classify the extracted features. With the feature level and decision level fusion methods, the whole recognition accuracy was improved more than 10% compared to the results based on gesture emotion recognition. Soleymani, Pantic, and Pun (2012) proposed a user-independent emotion recognition method by using electroencephalogram (EEG) signals, eye gaze data and users' bodily responses to videos (Soleymani, Pantic, & Pun, 2012) with the goal of recovering affective tags for videos. In their paper, Soleymani, Pantic, and Pun (2012) used a two-dimensional emotion model (the valence-arousal emotional space). The limitation of this work is that only a small video data set was used in the experiment.

This dissertation focuses on multi-sensory emotion recognition. Our goal is to design a framework for emotion recognition from speech via the microphone and facial expression via the built-in webcam or other camera device and then fuse the information of these two parts to improve the accuracy of emotion recognition. The results of our experiments show that this kind of method can achieve higher accuracy than using speech emotion recognition or facial expression recognition alone.

Emotion Models

There are plenty of challenges in the research of emotion recognition, such as what kind of emotions should be conceived to recognize and where to observe emotions since the information about emotion can be found in facial expressions, body gestures, speeches, or conversation. Actually, human beings are the experts in emotions since we use emotions all the time in our daily lives. We can also name the emotion we expressed to others and tell the emotion states the people we interacted with. However, it is hard to

describe emotions by a computer automatically. Moreover, to classify emotions by some rules is even harder. Thus, the models of emotions are significant.

Psychologists have done plenty of detailed research on emotions, and they have proposed multiple models or theories for the description of emotions. Generally, emotions can be divided into three categories: the basic emotion model, the dimensional model, and the componential appraisal model (Grandjean, Sander, & Scherer, 2008).

The basic emotion model

This category is based on the research of emotion theory from Darwin (Darwin, 1998) interpreted by Tomkins (Tomkins, 1962; Tomkins, 1963). Ekman represented the theoretical proposals of the basic emotion model in Ekman (1992) and Ekman (1999). According to the experiments of Ekman (1992), judging the static images with facial expressions of human, there are six basic emotions that can be recognized universally. These emotions are happiness, sadness, surprise, fear, anger, and disgust. From his research higher level emotions can be combined from six basic emotions. The higher level emotion recognition is not in the research scope of this dissertation, since most of recent emotion recognition researches focus on the basic emotion model such as the six basic emotions represented by Ekman (Ekman, 1992). Figure 1 is an example of the facial expressions of the six basic emotions.

Figure 1. The six basic emotional states (Kanade, Cohn, & Tian, 2000).

Plutchik coincided with Ekman's theory and developed the "wheel of emotions" (Plutchik, 1980). In Figure 2, the Plutchik's wheel, there are 8 emotions are arranged in opposite pairs (joy and sadness; anger and fear; disgust and acceptance; surprise and anticipation) with the strength of the emotions described in distinct colors. According to Plutchik's research, human beings cannot experience opposite emotions at the same time. Complex emotions, which could arise from a cultural condition or association with basic emotion, can be formed by just modifying some basic emotions. Though some researchers have proposed a different number of basic emotions which can range from 2 to 18 (Ortony & Turner, 1990; Wierzbicka, 1992), Ekman's theory of the six basic universal emotions is the most acceptable one in the research of emotion recognition areas.

Figure 2. Plutchik's wheel of emotions (Plutchik, 1980).

The concept of basic emotions is easy to accept; however, the validity is questioned. In Ortony and Turner (1990), the authors argued that the view of existing basic emotions (for example, anger, fear, happiness and so on) can build or explain all other emotions (contempt) is questioned, and the expression of emotions is not the same as the emotions themselves. For example, specific facial expressions that are recognized around the world and seem universal are not linked to emotions, but rather to certain conditions that also elicit emotions. Furthermore, it is often not enough to describe the emotions of daily life by the mixing of basic emotions. A real emotion may be between joy and surprise, shadowed by anticipation, and the intensity of all the components is also important. Ekman also reported that there is some confusion from the judgment study of the six basic emotions. For example, anger and disgust, as well as fear and surprise, are commonly confused. Surprise is also confused with the emotion of interest.

The Dimensional model

Emotions can also be represented in a dimensional framework where emotions can be mapped by two or three variables. According to this approach, the emotional states are not independent from each other. Variables for a two dimensional space are usually valence and arousal. For a third dimension energy or control is usually included. The valence dimension usually represents the positive or negative degree of the emotion, and the range is from uncomfortable feelings to comfortable feelings. The arousal dimension represents how excited the emotion is, and it ranges from low to high. While the energy or control dimension represents the degree of the energy or control over the emotion. These variables enable a more accurate description of emotions, since multiple aspects of emotions are used simultaneously in a continuous range of values. Basic emotions can still be represented in a dimensional model as a point or an area. As shown in Figure 3, according to the differences of valence and arousal, emotion can be represented in the two dimensional space. For example, happiness can be elicited by a beautiful sunset or a smiling baby (the yellow rectangle) in the arousal and valence dimensions. There are some studies on using dimensional models for emotion recognition (Grimm, Mower, Kroschel, & Narayanan, 2006; Wöllmer, et al., 2008), and usually only two dimensional models (valence-arousal model) are used.

The advantage of the dimensional model is that it is very intuitive to represent emotions on some continuous scale. However, the reduction of emotion differentiation to two or three dimensions will lose some information (Ekman, 1982; Tomkins, 1962). Moreover, some of the basic emotions proposed by Ekman, such as happiness or sadness, are easy to recognize in the dimensional models. However, some other emotions such as

anger and disgust are hard to distinguish, and some of the emotion cannot even be described. Take Figure 3 for example, surprise is missing in the two-dimensional model.

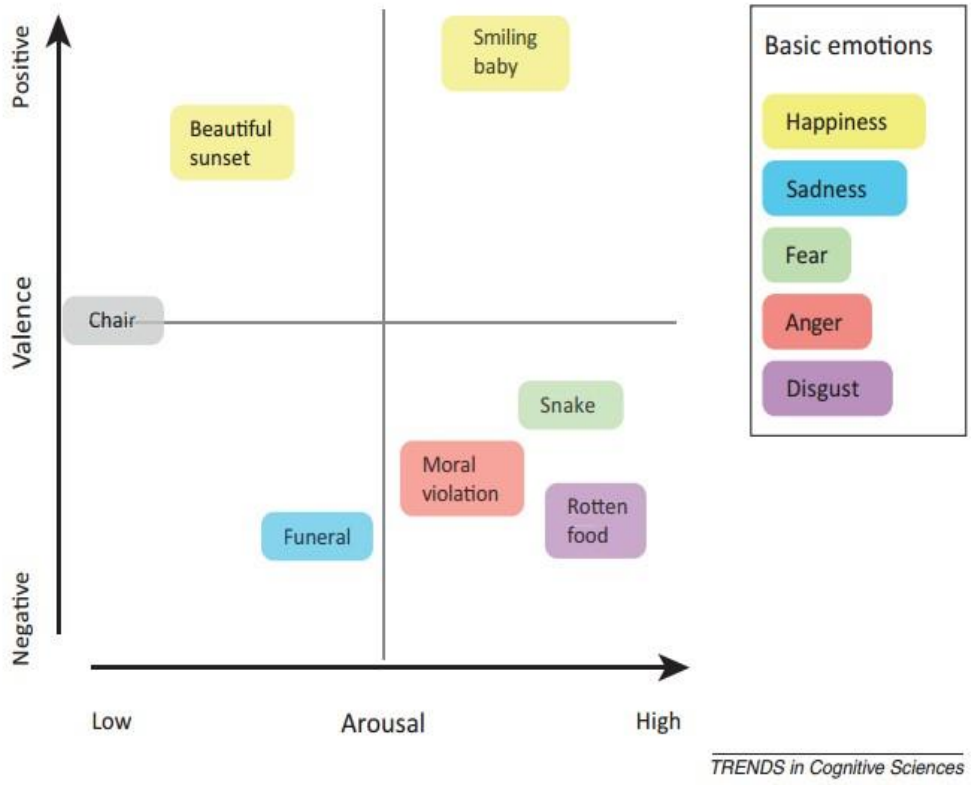

Figure 3. Distinctive and dimensional emotion model (Hamann, 2012).

The componential appraisal model

The componential appraisal model concerns identifying emotion according to the interpretation of events which cause emotions outside. This model can be seen as an extension of the dimensional model which proposed by Scherer (Scherer, Schorr, & Johnstone, 2001). It views emotion as a dynamic episode involving a process of continuous change in all relative components such as cognition, motivation, physiological reactions, motor expressions, and feelings.

In the componential appraisal model, emotions are defined as complex, multi-componential, dynamic process, and there is no limitation on the numbers and the

dimensional space of emotions. This model focuses on the changing of emotional states, predicts, and provides multi types of appraisal patterns. The richness of emotion differentiation of this model allows researchers to model individual differences and emotional disorders (Scherer et al., 2001). However, the measurement of emotional states changing in this model is complex and unsophisticated. Thus, it is still an open area for emotion recognition with the appraisal model.

The basic emotion model mentioned in this section can simply represent part of emotions and describe more or less for psychological purposes. With the current state the basic emotions or discrete emotional classes are widely used in automatic emotion recognition because a higher number of dimensions provided by other models are not reliable to estimate. Thus, in our research the basic emotion model is chosen in the design of the multi-sensory emotion recognition system.

Outline of the Dissertation

The remainder of this dissertation is organized as follows:

Chapter II introduces the background, related works on speech emotion recognitions, and emotional speech databases. While the issues on how to achieve speech emotion recognition with OpenSMILE are also discussed.

Chapter III includes the background and related works about the facial expression recognition. The general procedures—face detection, face tracking, feature extraction, and classification—are discussed in detail.

Chapter IV presents our framework of multi-sensory emotion recognition. There are mainly three parts that are concerned with the speech emotion recognizer, facial expression recognizer, and information fusion. The speech part was realized by the

software OpenSMILE. There are two separate concerns for facial expression recognition. One is used to process the static images, and the other one is utilized to process facial emotion recognition from videos or live camera inputs. In this system, we mainly integrate video streams to the multi-sensory emotion recognition framework. Information fusion is used to integrate the results from speech and facial emotion recognition. The weighted fusion techniques are presented in this chapter.

Chapter V talks about the testing we have done on the proposed system including choosing the base model of the speech recognizer, testing of speech and facial emotion recognition, results of information fusion, as well as the testing on static image for facial expression recognition.

Finally, in Chapter VI, some conclusion marks, and future work are summarized.

CHAPTER II

SPEECH EMOTION RECOGNITION

Introduction

Information about emotion is encoded in many aspects of speech such as what is said in a speech and the way it is said. The same sentence can be uttered in different ways to convey totally different emotions, while different people may prefer different words to describe exactly the same thing. Generally, there are two major types of messages (Johnstone & Scherer, 2000) in speech: the explicit (linguistic) message and the implicit (paralinguistic) message.

Explicit messages are the contents of what was spoken, while the implicit messages are the way how the contents were spoken. In speech emotion recognition, if one only considers the content, what was spoken, important information may be missed from the speaker's utterance, and even the content will be misunderstood. Furthermore, the linguistic content and emotion are language dependent, and a general rule is hard to conclude (Ortony & Turner, 1990). Our experiments on choosing the base model for speech emotion recognition also demonstrate that emotion recognition from speech is language dependent.

Implicit messages are usually the acoustic and prosodic features that can be used to infer the emotional states of a person when speaking. For example, the pitch feature, one of the most reliable features, can be seen as the index of arousal (Johnstone & Scherer, 2000). Currently, there are many approaches that have been done by researchers on detecting emotion from the implicit messages. However, the interpreting of paralinguistic features has not been fully discovered. Some ambiguities regarding how

different acoustic features affect different emotional states still exist. The advantages for recognizing emotion from audio information are easy to realize, low cost, and multiple available databases. The disadvantages are language dependent, lower recognition accuracy when compared to facial expression and ambiguity about the implicit messages.

Currently, emotion recognition from speech grows to an important research area. Several approaches to recognize emotions from speech have been proposed that are capable to detect emotion from speech by acoustic or linguistic information. So far, there are several off-line approaches for speech emotion recognition, and on-line speech emotion recognition has been achieved scarcely (Vogt, AndrÃl', & Wagner, 2008).

Applications

The recognition of emotion in speech may dramatically improve human computer interaction and can be used in a verity of applications. Call-Center Application (Petrushin, 1999) is an area where emotion recognition is widely used to detect the emotional state in telephone call center conversations and provides feedback to the operator or supervisor for monitoring purposes. Also, it can help to store calls and voice mail messages according to the callers' emotions or even transfer the call to the operator or supervisor if necessary. Another widely used application is the stress detection of emotional speech. Such a system can be used to monitor the stress level when a person is driving a car, making a phone call, or in a face-to-face interview. It also may be used to monitor a person's mental and psychological state to help the psychologist determine a patient's mental and physical health. This kind of application can also be used in lie detection. Furthermore, emotion recognition can be used to design friendly interfaces which can adapt them to the user's emotional states in intelligent human computer interaction. For

example, this can improve the user's gaming experience when playing a video game and makes the game more interesting and intractable.

Challenges

As mentioned in the previous chapter, emotion recognition from speech is difficult because of several events and challenges. First of all, human beings' emotions are extremely varied, various and ambiguous. It is frequently very hard for humans themselves to recognize others' emotions correctly, since each individual has his/her own manifestations of different emotions. This makes picking out a person's emotion via a machine or a computer even harder. The full scope and variety of emotions make the emotion recognition much harder. It is inconceivable to count all the spectrum of emotions. Some subtle emotions like pride, appreciation, etc. may not yet be given voice. Most emotional recognition systems can simply make out a little set of emotions, for example the six basic emotions. Another question for emotion recognition from speech is what data is the best. Since there are different ways to show emotions, what data is chosen to best represent different emotions on the spot? There is a great amount of information that can be collected, and many characteristics can be drawn out. So far, researchers have not been able to zero down on the feature-sets which are better than others for speech emotion recognition. Finally, the lack of standardized data sets makes the comparison of recognition results more difficult. Since researchers use different data sets and different features to perform emotion recognition, to compare the resolutions and achieve a consensus on the practice is hard. How to collect the data for emotion recognition is also a major problem. The recognition rate is high in experiments while testing in the actual world the similar result is harder to achieve. The high quality of

records held for training purposes have significantly better performance than what we can collect in our daily lives during the experiments. Many organizations such as the IEEE Interspeech are now making efforts to estabilsh standards for emotion recognition (Bozkurt, Erzin, Erdem, & Erdem, 2010).

Related Works

As we described previously, the research of emotion has a long history in psychology. In this section several excellent research works on emotion recognition from speech will be briefly summarized.

In the final few decades, a number of researches on affective computing have been nominated. In history, the first practical pilot work on automatic emotion recognition was conducted in 1996 by Dellaert, Polzin, and Waibel. In their work, several statistical pattern recognition techniques to classify utterances according to the emotional content were proposed. They have collected over 1000 utterances from different speakers in four different, acted emotional states: happiness, sadness, fear, and anger. In their work, only the pitch features were taken out from the utterances for classification purposes. A novel pattern recognition technique named majority voting of subspace specialists was proposed. By applying this method, the accuracy of the carrying out classification may achieve 80% accuracy. Since the book *Affective computing* (Picard, 1997), researchers have assumed that emotional intelligence can be used to help us solve emotion recognition problems. An excellent overview of current trends in emotional speech recognition is represented by Ververidis and Kotropoulos (2006). Vogt (2010) is a comprehensive dissertation that proposed real time identification of emotion from spoken language. In their work, a real-time speech emotion recognition system called EMO

VOICE was implemented. The Naïve Bayes and Support Vector Machine (SVM) classification algorithms were used in EMO VOICE to classify emotions.

In speech emotion recognition, feature sets are very important for training the classifier and doing classification. These features can be either local or global and linguistic or acoustic. In Schuller et al. (2007), over 4000 features were collected together and grouped into 12 low level descriptor types and 6 functional types. In Baltiner et al. (2011), an extensive feature selection method, Sequential Forward Floating Search was proposed and performed on 7 acoustic and 4 linguistic types of features. Different acoustic feature sets were ranked by their method. They showed that the performance of all acoustic and linguistic features is comparable, and then showed that the combination of both features can improve the recognition accuracy. Vogt and Andre (2005) compared the feature sets from 1000 features derived from pitch energy and the MFCC time series of acted and realistic emotional speech. They concluded that there are significant differences in the feature sets of different type of databases. In Schuller, Muller, Lang, and Rigoll (2005), the acoustic features and linguistic features were integrated to perform a more robust estimation. The results demonstrated that by fusing the acoustic and linguistic features in a vector, the emotion recognition performance was improved remarkably for existing databases. An excellent study of the progress of speech emotion recognition in the past fifteen years is provided in Schuller, Batliner, Steidl, and Seppi (2011) which introduced "where we are today, how we got there and what this can reveal us on where to go next and how we could arrive there." The authors also pointed out that obtaining more realistic data is still the most significant challenge in the future of speech emotion recognition. New features can be considered on speech emotion recognition such

as gender and age. The gender-dependent emotion recognizer in Vogt and Andre (2006) performed better than the gender-independent ones. In their paper a gender detection system was used with the accuracy of 90%. The combination of gender detection and emotion recognition system can enhance the overall recognition accuracy of 2%-4%.

Factors of Speech Emotion Recognition

From the related works on speech emotion recognition, we have noticed that for speech emotion recognition there are usually 3 main steps: signal processing, feature exaction and selection, as well as classification (Walker et al., 2004). In the first step, the acoustic signal is processed and segmented into smaller units. In the second step, the features of these small units were extracted and selected to represent different emotions. Finally, in the classification step, models are trained with emotion labels and then used to predict the emotion of a new instance. In this section some of the basics of speech signal concepts will be introduced. Then the details of the three major steps will be described. Finally, the feature we used in this project and some widely used speech databases will be given.

Basic Signal Concepts

In the signal processing step, the oral speech signal will be processed in a digital representation which includes sampling, filtering, and digitization of the recorded signal.

In this section some of the main concepts and terms of speech processing will be introduced as follows:

Frequency. It is the number of times a repeating event occurs in 1s and can be measured in hertz (Hz).

Amplitude. It is the value on the vertical axis which shows the air pressure as in Figure 3.

Pitch. Pitch is the sensation of a frequency. Thus, a high pitch sound corresponds to a high frequency sound wave, and a lower pitch means a lower frequency sound wave.

Loudness. Loudness corresponds to the energy of the speech signal, and it is related to the square of the amplitude.

Spectra. The complex waveform can be thought to be made of simpler waves at different frequencies. The spectra are the description of different frequency components of the wave and can be calculated by using the Fourier Transform.

Linear Predictive Coding (LPC). LPC is a popular technique to encode a speech signal. It can smooth out spectra and make finding the spectral peaks easily.

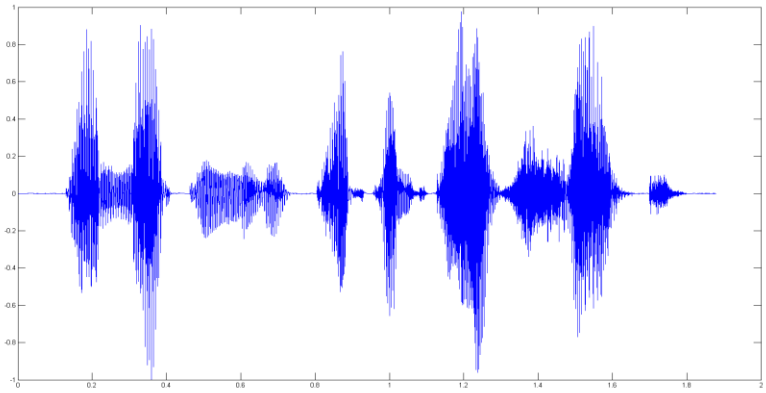

Figure 4. Analog signals of a recorded speech.

Spectrogram. A spectrogram is a visual representation of the spectrum of frequencies in a sound. It shows frequency changing over time. As shown in Figure 4, the x-axis is the time, and the y-axis is amplitude color-mapped as the blue bars.

Formants. It is the spectral peaks of the sound spectrum which is known to model spoken content, especially lower ones.

Speech Data Preprocessing

Now, let us talk about the data preprocessing step. In order to use the speech signal in the emotion recognition system, the analog signal which is only the representation and compression of air particles should be transformed into a computer understandable form.

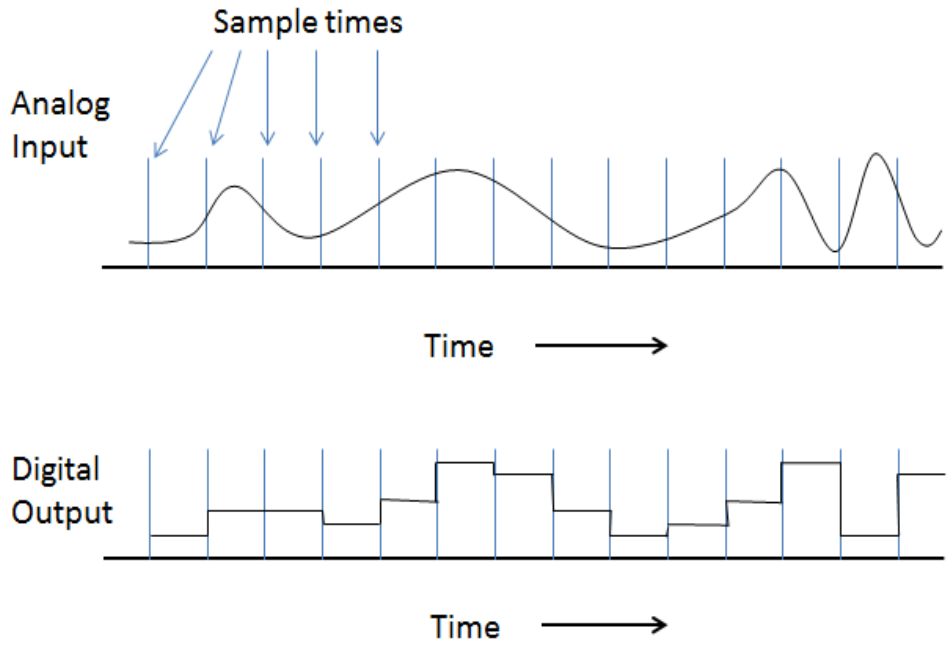

Figure 5. Digitizing of speech signal.

The signal should first be sampled, and its amplitude is measured at fixed time intervals. In order to capture all the contents in the speech signal, the sampling rate should obey the Nyquist Theorem. Then, the sampled signal is quantized. The real value of the amplitude will be converted to an integer value to enhance the high efficient

storage. Finally, the signal is digitized and suitable for analysis by computer. The signal transformation process can be seen in Figure 5.

Table 1

Different emotion units of speech

Unit	**Features based on**
phonemes	phonemes
words	Words Words in context All phonemes Vowels Voiced consonants
Utterances and turns	Utterances All words Central words
Fixed length intervals	Fixed length intervals
Relative length intervals	Relative length intervals

In order to represent emotions, the digitized signal must be segmented into smaller units. The segmentations should be long enough to calculate the features reliably and should be short enough to guarantee the stable acoustic property of the emotion (Walker et al., 2004). As shown in Schuller et al. (2011), the segmentation should be either technical (frame, time slices) or meaningful (like words, utterances, phrases). Finding the right unit is important in the final classification. However, there are no rules for finding the right unit, since it depends on the different cases and different data sets. In Batliner, Steidl, Seppi, and Schuller (2010), the word-based segmentation was used

though it is time consuming and expensive. A detailed overview of different segmentation units can be found in Table 1 (Vogt, 2010) where there are 3 units – phonemes, words, as well as utterances and turns in two different intervals.

Feature Extraction and Selection

Feature extraction and selection are significant for automatic emotion classification. However, there are no standardized feature sets for the best emotion recognition. Different features are used for different researches. Usually, the pitch and energy related features are used. As we have described previously, there are acoustic and linguistic features. In this dissertation, only the acoustic features are used to build a content independent system. There are two kinds of acoustic features: global features and short term features (Vogt, 2010). They are also called functionals and low level descriptors (LLDs), respectively (Schuller et al., 2007). The short term features are computed on segments that are in very short time intervals (10-30ms) with a constant length. A summary of LLDs or short term features can be seen in Table 2. LLDs or short term features contain timing information about emotions, can be used in dynamic classifiers, for example the Hidden Markov Models (HMMs). While the global statics features or functionals are calculated by applying statistical functions on the preprocessed segments. Various functions can be used as the statistical functions, such as mean, maximum, minimum, percentiles, etc. The global statistic features are not time sensitive, so we can use static classifiers (such as Support Vector Machines [SVMs]) to do the classification of emotions. A detailed comparison about different features can be found in (Baltiner et al., 2011).

Table 2

A summary of LLDs/short term features

Type of feature	**Descriptions**
Duration	Temporal aspects, for example F0 value (measured in milliseconds)
Energy	Describe the intensity or amplitude
Pitch	Related to the tone
Spectrum	Good representation of human characteristics, including formants, pectral slope, mean, center gravity, etc.
Cepstrum	Hold frequency as well as timing information. The most popular is MFCC
Voice quality	Includes jitter, shimmer, and Harmonic to Noise Ratio (HNR). Nonlinear operators such as Teager Energy Operator
Wavelet	Similar to MFCC but more localized. Contain more information

After the extraction of features, feature selection is the next problem to be solved. Feature selection is used to reduce the dimensions of the feature space. To obtain a better feature sets many feature selection methods were introduced to this area, such as Sequential Forward Selection (SFS) algorithm or Sequential Forward Floating Search (SFFS) algorithm with a k-Nearest-Neighbor or Support Vector Machine classifier (Lee & Narayanan, 2005; Schuller et al., 2005; Ververidis & Kotropoulos, 2006). Principal Component Analysis (PCA) can be used to reduce the feature dimensions (Fernandez & Picard, 2005; Lee & Narayanan, 2005). The Enhanced Lipschitz Embedding (ELE) is another method for feature space reduction based on geodesic distance estimation (You, Chen, Bu, Liu, & Tao, 2007). Emotion recognition from speech with feature selection might achieve the same or a little bit less accuracy than that of without feature selection, but much more efficient and less time consuming.

Classification

Generally, there are two kinds of classifiers: the static ones and the dynamic ones. The static classifiers mainly work on the global statistic features and assign class labels while the dynamic ones work on short term features or LLDs. The data dimensionality and the amount of training data are some of the factors when choosing a suitable classifier.

For static classifiers, many famous classification algorithms have been applied in this area, including Maximum likelihood Bayes (MLB), Neural Networks (Petrushin, 2002; Xiao, Dellandrea, Dou, & Chen, 2007), Decision Trees (Cichosz & Slot, 2007; Sidorova, 2009), Bayesian classifiers (Barra-Chicote et al., 2009; Cho, Kato, & Itoh, 2007; Lugger & Yang, 2008; Mower, Mataric, & Narayanan, 2009; Planet, Iriondo, Socoró, Monzo, & Adell, 2009), Gaussian Mixture Models (GMM) (Clavel, Devillers, Richard, Vasilescu, & Ehrette, 2007; Lugger & Yang, 2008; Neiberg & Elenius, 2008; Sethu, Ambikairajah, & Epps, 2009; Schuller, Steidl, & Batliner, 2009), Linear Discriminant Analysis (LDA) (Batliner, Zeißler, Frank, Adelhardt, Shi, & Nöth, 2003; Lee & Narayanan, 2005; Neiberg & Elenius, 2008), k-Nearest Neighbors (k-NN) (Hassan & Damper, 2009), SVMs (Casale, Russo, Scebba, & Serrano, 2008; Oudeyer, 2003; Vlasenko, Schuller, Mengistu, Rigoll, & Wendemuth, 2008; Wu, Falk, & Chan, 2008; Zhou, Wang, Yang, & Chen, 2006), and Random Forest (RF). Only some of these classifiers are suitable for emotion recognition. For large-scale data sets, a data mining software Weka (Hall et al., 2009) has been widely used in the experiments of classifiers comparison. Dellaert et al.(1996) has compared the performance of the MLBs, k-NNs, and kernel regression classifiers. Their results show that k-NNs can give better

performance than the other classifiers. However, k-NNs are not suitable for high dimensional input data. Many researches (Casale et al., 2008; Oudeyer, 2003; Schuller et al., 2005) show that SVMs have a better recognition rate than that of the Decision Trees, the Naïve Bayes, as well as the k-NNs. Thus, the SVMs now are the most popular and successful static classification methods for using in speech emotion recognition area.

When it comes to dynamic classifiers the most popular algorithm is the Hidden Markov Models (HMMs) (Fernandez & Picard, 2003; Kwon, Chan, Hao, & Lee, 2003; Lee, Yildirim, Bulut, & Kazemzadeh, 2004; Wagner, Vogt, & Andre, 2007). For classifying short utterances HMMs have the best performances among HMMs, SVM, and LDA in Kwon et al. (2003). In Wagner et al. (2007), a distinct research of different HMM designs has been compared. The best design shows to be database and task dependent, and HMMs also perform better on short units than longer ones. Currently, HMMs are also successfully used in Automatic Speech Recognition (ASR).

Recently a hybrid model of HMM and SVM classification method was used in Aastha (2013). In their work, HMM was used in system training, while the SVM classifier is used in system testing. Compared to the SVM classifier, the accuracy improvement of this hybrid model is 4%. So far, the static classifiers have more applications than the dynamic ones. However, whether to use a static or a dynamic classifier to do the classification is still an open area.

Useful Features

As we said in the previous section, there are two kinds of features to use the LLDs and the global statistic features calculated by different functionals on these LLDs. In this dissertation, the INTERSPEECH 2009 Emotion Challenge feature set (Schuller et al.,

2009) and another feature set are used. The first feature set contains 384 features by applying 12 functionals to 16 LLDs. The second feature set contains 988 features by applying 19 functionals to 26 LLDs.

Most of the LLDs used in the above two feature sets are as follows:

pcm_RMSenergy. It means the root-mean-square signal frame energy. Actually, it is the square root of the average sum of the squares of the amplitude of the signal samples.

mfcc. Mel-Frequency Cepstral coefficients 1–12. These features are strongly dependent on the spoken content and the homographic transform with equidistant ban-pass-filters on the Mel-scale. They can benefit speech processing tasks practically by emphasizing changes or periodicity in the spectrum. Also, they are very robust against noise. Figure 6 is a block diagram for calculating MFCCs.

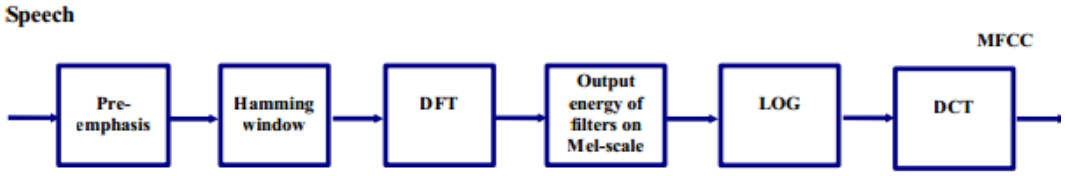

Figure 6. MFCC calculation.

pcm_intensity. The amplitude of the air pressure.

LspFreq. Line spectral pair frequencies coefficients 1–8 are used to represent linear prediction coefficients for transmission over a channel.

pcm_zcr. It is the Zero-crossing rate of time signal (frame-based). The zcr is the rate of sign-changes along a signal. This feature is widely used in speech recognition and music information retrieval. The definition of zcr is as follows:

$$zcr = \frac{1}{\mathrm{T}-1}\sum_{t=1}^{T-1} \Pi\{S_t S_{t-1} < 0\} \qquad (1)$$

where *S* is the length of a signal, *T* and the indicator function Π{*X*} is 1 the argument *X* is true and 0 otherwise.

voiceProb. This is the voicing probability computed from the Auto Correlation Function (ACF). It is related to the pitch and given by the following equation:

$$10 \cdot \log \frac{\mathrm{ACF}\ (\mathrm{T}_0)}{\mathrm{ACF}\ (\mathrm{T})} \tag{2}$$

where T_0 is the pitch period.

F0. It is the fundamental frequency computed from the spectrum. It is the acoustic equivalent to the perceptual unit pitch and measured in Hz. It can help to model intervals, characterize points, or contours.

Some of the functionals are given in the following:

max. This is the maximum value of the particular LLD.

min. The minimum value of the given LLD.

range. The values are between the max and min.

maxPos. The absolute position of the maximum value (in frames).

minPos. Similar to the above, it is the absolute position of the minimum value in frames.

amean. This is the arithmetic mean of the contour.

linregc1. This is the slope β of a linear approximation of the LLD. Suppose n data points $\{x_i, y_i\}$ are given and $i = 1, 2, \ldots, n$. The goal is to find a straight line according to the points

$$y = \alpha + \beta x \tag{3}$$

by the least-squares approach. Then we can get β as:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (4)$$

linregc2. This is the offset α of a linear approximation of the LLD. We can get the value of α by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad (5)$$

linregerrQ. This is the quadratic error computed as the difference of the linear approximation and the actual contour which is also called Residual Sum of Squares (RSS). The RSS for Equation (3) is:

$$RSS = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2 \quad (6)$$

stddv. This is the standard deviation of the values in the contour. It is the square root of the second central moment named variance. n^{th} order central moment of a continuous uni-variate probability distribution can be calculated in Equation (7)

$$\mu_n = \int_{-\infty}^{+\infty}(x-\mu)^n f(x)dx \quad (7)$$

where μ is the mean of the distribution and $f(x)$ is the probability density function.

skewness. Refers to Equation 7. It is the 3rd order central moment.

kurtosis. Refers to Equation 7. This is the 4th order moment.

Databases of Emotional Speech

Databases with emotional speech are important not only for psychological research, but also for the emotion recognition. A detailed description of 32 emotional speech databases was reviewed in Vereveridis and Kotropoulo (2003). In this section we only introduce some of them.

Generally, databases of acted, induced, or completely spontaneous emotions are used in research. The complexity increases with the naturalness of the speech databases. Thus, the best results are usually obtained with the acted emotion databases because of the strong emotion in those databases. The most famous acted databases are the Danish Emotional Speech corpus (DES) (Engberg & Hansen, 1996) and the Berlin database of emotional speech (Emo-DB) (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005). In DES there are 5 basic emotions recorded by 4 persons while in Emo-DB there are 7 basic emotions acted by 10 persons. In this work, we use the Emo-DB database to construct one of the base models for our speech emotion recognition system. The SmartKom corpus (Schiel, Steininger, & Turk, 2002) and the German Aibo emotion corpus (Batliner et al., 2004) are induced databases recorded in a lab to fulfill certain tasks. While the call center communication in Devillers, Vidrascu, and Lamel (2005) obtained from live recordings can be treated as the realistic database.

The interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an acted multimodal and multi-speaker database collected by researchers at the University of Southern California (Busso et al., 2008). It can be downloaded freely with permission from the website (IEMOCAP DATABASE, 2014). This database contains approximately 12 hours audiovisual data, including video, voice communication, facial motion capture, text transcriptions. Though it is an acted database, the elicited emotional expressions are mainly done by the actors. The motion capture information, the interactive setting to elicit authentic emotions, and the large size of the database make it valuable for the study and research in modeling multimodal emotion recognition.

In human computer interaction, the more realistic the data the harder it is to recognize an emotion from it. Figure 7 shows the difficulty of emotion recognition with different types of databases.

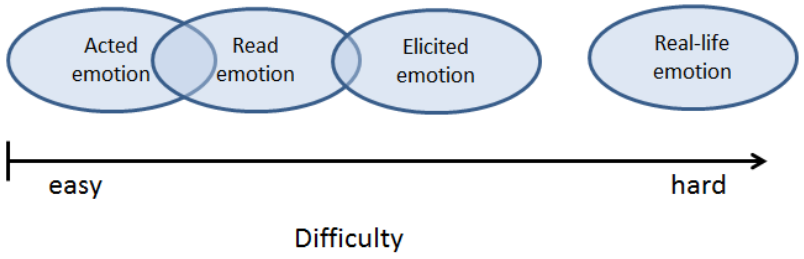

Figure 7. Difficulty of speech emotion recognition with different types of databases.

CHAPTER III

FACIAL EMOTION RECOGNITION

Introduction

Facial expressions are fundamental for social communication because they hold significant clues about emotions directly. Facial movements have several roles in the interaction and communication between human beings. Visible facial movements are used to enhance and influence the emotion from speech. Moreover, facial actions are activated for a short time when the emotion passes. Thus, detecting the facial expression is an intuitive way to recognize emotion.

The cross cultural study of Ekman (Ekman, 1999) shows that some emotion expressions are universal for human beings regardless of their race and region. These emotions are happiness, sadness, anger, fear, disgust, and surprise. Each of the six basic emotions corresponds to a unique facial expression, and other emotional expressions may be culturally variable. Table 3 gives the description of facial expressions of the basic emotions.

For emotion recognition systems, facial expression analysis is considered to be a major indicator of a human affective state. Automatic emotion recognition from facial expression is inherently a multidisciplinary enterprise involving different research fields (Friesen, Ekman, & Hager, 2002) including psychology, computer vision, feature data fusion, and machine learning. There are two main streams in current facial emotion recognition: the facial affect recognition and the facial muscle action detection. They came from two dominant facial expression analysis methods in psychological research: the message judgment method and the sign judgment method. The message judgment

method is used to understand what underlies a displayed facial expression; while the sign judgment method is used to describe the shown behavior outside.

Table 3

Facial Expression Description of six Basic Emotions (Sumpeno et al., 2011).

No	***Emotion Name***	***Description of Facial expressions***
1	Happiness	The eyebrows are relaxed. The mouth is open and the mouth corners upturned.
2	Sad	The inner eyebrows are bent upward. The eyes are slightly closed. The mouth is usually relaxed.
3	Fear	The eyebrows are raised ad pulled together. The inner eyebrows are bent upward. The eyes are open and tense.
4	Anger	The inner eyebrows are pulled downward and together. The eyes are wide open. The lips are tightly closed or opened to expose the teeth.
5	Surprise	The eyebrows are raised. The upper eyelids and the eyes are wide open. The mouth is opened.
6	Disgust	The eyebrows and eyelids are relaxed. The upper lip is raised and curled, often asymmetrically.

The facial action coding system (FACS) proposed by Ekman and Friesen (Ekman & Friesen, 1971) is a commonly used vision-based tool to code human facial expression movements by their visual aspect on the face which belongs to the sign judgment method. The FACS enables facial expression analysis through standardized coding of changes in facial motion in terms of atomic facial actions named Action Units (AUs). The FACS can decompose the facial muscular actions and key out the facial expression in AUs. Then, the AUs can be applied for any high-level decision making process including basic

emotion recognition, various affective state recognition, and other complex psychological states. This system codes facial expressions manually by following a set of specific rules. So far, Ekman's basic facial emotion model and the FACS are the mainly used methods in vision-based facial emotion recognition. However, the inputs for the FACS are static images of facial expressions which result in a very time consuming process.

Ekman's (1997) pioneer work inspired many other researchers to analyze facial expressions via images and videos. The facial expressions were categorized by the facial feature tracking and facial movements measurement. Some survey works (Fasel & Luettin, 2003; Pantic & Rothkrantz, 2000; Pantic & Rothkrantz, 2003) give a deep review of much research done in the field of automatic facial expression recognition. In general, there are four steps for recognizing emotion from facial expression: face detection, face tracking, feature extraction, and classification. Few facial emotion recognition systems can deal with both static images and image sequences. For example, the survey published by Pantic and Rothkrantz emphasized how to automatically analyze the facial expression (Pantic & Rothkrantz, 2000). The face was detected using watershed segmentation with markers extracted on the Hue Saturation Value (HSV) color model algorithm. A point-base face model is used, and the features are defined as some geometric relationships between the facial points or the image intensity in a small region defined relative to the facial points. Multiple feature detectors are used for each facial feature localization and model feature extraction. The reported recognition rate is 86% by using the rule-based classification method. In Lyons, Budynek, Plante, and Akamatsu (2000), a set of multi-scale, multi-orientation Gabor filters were used to transform an image. Then an elastic graph matching method is used to obtain a registered grid. The grid was sampled and

combined into one vector as features. When testing with a database of 193 images, 75% accuracy was achieved. By detecting the eye and lip position using low-pass filtering and edge detection method in (De Silva & Hui, 2003), an average recognition rate of 60% was achieved. A bimodal system of facial emotion recognition has been proposed by Wang and Guan (2008). An HSV color model was used to detect the face of the environment, and Gabor wavelet features were used to represent the facial expressions. The overall recognition rate is 82.84% by using multi-classifiers.

Though there are many different methods in facial emotion recognition, the common steps are face detecting, face tracking, feature extracting, and classification. The output is the emotion recognition result in the preselected basic emotions. Basically, the facial emotion recognition design in this dissertation also follows these common steps. The details will be described in Chapter IV.

Related Works

A facial expression is a visible activation of a human's affective state, cognitive activity, temperament and personality, and psychopathology (Donato, Bartlett, Hager, Ekman, & Sejnowski, 1999). Facial expressions together with other gestures can reveal non-verbal communication hints in the face-to-face interactions, which can help the listener to understand more about the intended meaning of the speaker's speech. FACS proposed by Ekman and Freisen (1978) can describe the distinguishable visual facial movements. Using the FACS, the parameters of action are designed to classify human emotions. From Mehrabian's research (1971), facial expression provides 55% of the effect of a message. The vocal part contributes 38% while the verbal part only contributes

7%. Facial expression applications are widely used, including image understanding, psychological studies, medicine, and image compression (Ostermann, 1998).

Since facial expressions include a consequence of information about a person, they play an important role in the human computer interaction. Automatic emotion recognition from facial expression may act as a component of the natural human machine interface (Dam, 2000). This kind of interface can provide services requiring a good understanding of the emotional state of the user. For example, using this interface in some robots to recognize human expression can contribute more intelligent robots (Bruce, 1992). Automatic facial expression analysis for behavioral science or medicine is another kind of application (Donato et al., 1999; Essa & Pentland, 1997).

Many works have been done with automatic emotion recognition from images or videos of facial expression since 1990. In Mase (1991), optical flow (OF) was used by the author to estimate facial muscle movements in order to recognize facial expressions. It was the first time to use image processing techniques in facial expression recognition. A flexible shape and appearance model for locating facial features, coding and reconstruction, recovering pose, recognizing gender and facial expression, and identifying individuals with an image was used in Lanitis, Taylor, and Cootes (1995). Black and Yacoob (1995) used a collection of local parameterized OF models to track and recover rigid and non-rigid facial motions. The image motion parameters were then used in a rule-based classifier to recognize the six universal facial expressions. Eye blinking and other simple head motions can also be recognized by their system. In Yacoob and Davis (1996) the authors computed the OF and used similar rules to Black and Yacoob (1995) to classify the six basic emotions. OF regions on the face were

computed in Rosenblum, Yacoob, and Davis (1996), and a radial basis function network architecture was used on the human emotion detection system. OF flow processing was also used in Essa and Pentland (1997) as the basis to measure and classify facial emotions. In Otsuka and Ohya (1997), the OF algorithm was first used to calculate a velocity vector. Then, a two dimensional Fourier transform is used for the vector at the eyes and mouth regions. Finally, Hidden Markov Models (HMMs) are used to recognize facial expression from image sequence for multiple persons. In their experiments, the proposed system can recognize basic emotions in nearly real time.

Although there are many achievements for emotion recognition from facial expression, there are still many difficulties and limitations in emotion recognition due to the complexity of emotion expression, especially when in a conversation. The advantages of recognizing emotion from facial expression are a) it is the most intuitive and natural way to observe human beings' emotional states; b) there are plenty databases available for research. One disadvantage is that it is misleading sometimes since there is no context information. Another one is that the recognition results depend heavily on the quality of the image or video.

Recently, many efforts have been done on implementing emotion recognition systems using both facial expressions and acoustic information. De Silva and Pei (2000) proposed a bimodal emotion recognition system by combining audio and video information using a rule-based system. The authors described the use of statistical techniques and hidden Markov models (HMM) in the emotion recognition. The prosodic features in audio and maximum distances and velocities of video between six specific

facial points were used. The performance of emotion recognition increased when using both ways together.

Factors of Facial Emotion Recognition

Vision based emotion recognition is mainly focused on facial expressions because of the significance of face in emotional expression and perception. Many approaches have been done in this area. However, there are still many challenges in the facial emotion recognition area. As we know faces are non-rigid and have different color and pose. Some of the facial features are not common and not suitable for pattern recognition. Lighting of the background and the illumination conditions can also change the overall recognition rate of facial expression. The above problems make the emotion recognition from facial expression more complicated.

Currently, many facial expressions recognition approaches are based on a two dimensional (2D) sparse data, such as 2D static images or 2D video sequences. In this research, we also use the 2D method to do our emotion recognition from facial expressions. As we have introduced in the previous chapter, emotion recognition from speech needs three main steps, while for the facial emotion recognition, there are mainly four steps: face detection, face tracking, feature extraction and classification as shown in Figure 8. The input data for the recognition system can be video streams from a Web Camera or some recorded video clips, as well as some static facial images. After the 4 steps, the final decision of the facial emotion will be given as the output.

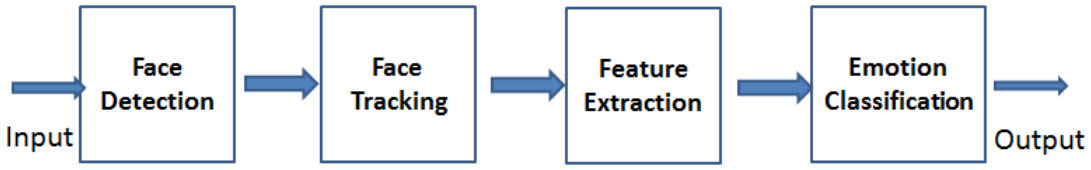

Figure 8. Facial emotion recognition.

Face Detection

Though there are several types of input to the facial emotion recognition system, images containing faces are still essential to intelligent vision-based human computer interaction. Face detection tries to identify all image regions which contain a face regardless of its 3D position, orientation, and lighting conditions from a given image. The challenge here is that faces are non-rigid and have a high degree of variability in size, shape, color, and texture. A variety of face detection techniques have been developed. In the Yang, Kriegman, and Ahuja (2002) survey, the authors classified the detection methods into four categories: Knowledge-based methods, Feature invariant approaches, Template matching methods, and Appearance-based methods.

Knowledge-based face detection methods are also called rule-based methods. These kinds of methods encode human knowledge of a typical face and are used mainly for face localization. Facial features will be extracted first from the input image, and then faces are identified based on the coded rules. The problem with this kind of method is that it is difficult to translate human knowledge in well-defined rules. Furthermore, it is difficult to extend this approach to detect faces in different poses. In Yang and Huang (1994), a hierarchical knowledge-based method was used by the authors to detect faces.

Compared to knowledge-based methods, the feature-based methods use the invariant face features to detect a face. These features are eyebrows, eyes, nose, mouth,

and hairline which are extracted by using edge detectors. Then, based on the extracted features a statistical model is built to verify the existence of a face. Feature-based methods are widely used by many researchers (Graf, Chen, Petajan, & Cosatto, 1995; Sirohey, 1993). The weakness for this kind of methods is that image features depends severely on the illumination, noise, and occlusion.

In template matching methods, a standard face pattern will be predefined with a function by experts. The correlation values for the face contour, eyes, nose, and mouth of a given image will be calculated with the standard patterns. Whether a face exists is dependent on these correlation values. This model is very easy to implement, but it is inefficient to deal with scale, pose, and shape for face detection. There are two types of templates in this kind of methods: the predefined templates (Sakai, Nagao, & Fujibayashi, 1969; Samal & Iyengar, 1995; Tsukamoto, Lee, & Tsuji, 1994) and deformable templates (Kwon & da Vitoria Lobo, 1994; Lanitis et al., 1995).

Finally, the models of appearance-based methods have been learned from a set of training images. They rely on techniques from statistical analysis and machine learning to find the relevant characteristics of face and non-face image. Many distribution models or discriminant functions are used to model the learned characteristics for face detection, including Eigenface (Turk & Pentland, 1991), Gaussian distribution (Sung & Poggio, 1998), Neural Network (Rowley, Baluja, & Kanade, 1998), SVM (Osuna, Freund, & Girosi, 1997), Naïve Bayes Classifier (Schneiderman & Kanade, 1998), HMM (Rajagopalan et al., 1998), and Information-Theoretical Approach (Colmenarez & Huang, 1997; Lew, 1996). According to the survey by Zhang and Zhang (2010), the

appearance-based methods are superior to the other method despite their computational load.

There are lots of challenges for face detection such as non-rigid structure, different illumination environment, size, shape, color, orientation as well as texture. Yang, Kriegman, and Ahuja (2002) pointed out that gestures, presence or absence components, facial expression, partial or full occlusions, face orientation, and lighting conditions are the basic challenges in face detection.

Viola and Jones (2001) have proposed a frontal face detection system in grayscale images using the Harr Feature-based Cascade Classifiers. It can be used in a real-time face detection system. Now it is available for all the researchers in the Open Source Computer Vision Library (OpenCV) tool (OpenCV, 2014). Their work has been extended to handle multi-pose faces using skin-color cues to reduce the computation time and decrease the false detection rate, since the skin color is a useful cue for face detection under different illumination environments. Currently, the Viola and Jones face detecting method is still widely used by researchers and also used in our research.

Face tracking

Face tracking is used in the video frames to pass over the detected face forward or backward of each frame. Usually, two methods can be used in face tracking. The first one uses the face detector as a face tracker running at all the frames. For instance, the Viola and Jones (2001) face detector can be used as a face tracker running at each frame. The second approach is to develop a face tracker apart from the face detector. Active Shape Models (ASM) (Cootes, Taylor, Cooper, & Graham, 1995) and Active Appearance Models (AAM) (Cootes, Edwards, & Taylor, 2001) based tracker are widely used.

A face shape should be reconstructed to fit the target face image in the face tracking. For the ASM based method, manually labeled training images are used (Cootes et al., 1995). This approach will first search the salient points to fit the model in the image and then update these points at each frame. It is also called Smart Snake. The AAM also uses a training phase, but it uses both shape and appearance information of the target image (Cootes et al., 2001). ASM is faster than the AAM because AAM uses all information of the image. The original algorithms were used for the grayscale images, but both of them can be extended to the color image.

ASM generation and fitting. Point distribution models (PDM) are important in modeling of shapes. The statistical information of the training images is used to extract the mean and variance of the shape. Landmarks on the boundary are used when describing an object. A transformation is used to align all the images in the training process including translation, scaling, and rotation. Each image in the training database has been co-aligned with this transformation. When the algorithm converges, the relationship between the original shape a and the mean shape $\bar{a}$ can be described as:

$$a = \bar{a} + T \times w,$$

where T is the matrix containing the eigenvectors of the shape in its columns, and w is the variation weighting matrix for each of the eigenvectors. The generated ASM model must fit the targeting object. After initialization, landmarks are moved along the search path, and the model boundary is fit to the target object boundary. For the detail, please refer to Sonka, Hlavac, and Boyle (2008).

AAM generation and fitting. In AAM generation, not only a shape model, but also an appearance model was generated. As we said previously, the ASM model is

generated only based on shape; while the AAM model is generated from both shape and appearance.

Feature Extraction

The extraction of facial features plays an important role in emotion recognition from facial expression. Among the facial features, eyes, nose, and lip are the most important features. Various approaches have been done to extract facial points (such as eyes and nose) from images and video sequence of faces. Also, feature extraction closely relates to the face detection, since in the face detection part some of the features have already been extracted. Here, we just simply introduce the concepts of feature extraction.

Generally, there are four kinds of methods for facial feature extraction: geometry-based feature extraction, template-based feature extraction, color segmentation-based feature extraction, and appearance-based feature extraction. Geometry-based approaches extract features using geometric information such as relative positions and sizes of the face components like mouth, eyes, nose, and eyebrow, which can cover the variation in the appearance of the facial expression. The template-based feature extraction matches facial components to previously designed template using the appropriate energy functional. The color segmentation method uses skin color to segment the face, and the non-skin color region will be viewed as a selection for eyes and mouth. This method is used in our approach of facial expression recognition. The appearance-based approach provides not only the simple components of the face, but also information about the texture of the face, such as wrinkles. To extract the feature vector, many methods can be used such as Gabor-wavelets and principal component analysis (PCA).

Emotion Classification

In the literature, there are many studies on classifiers of facial emotion recognition. For the static images, the facial expression will be classified based on the tracking result from each image. Bayesian network classifiers and Naïve Bayes classifiers are used generally to classify the facial emotions. While for the dynamic approach, the temporal pattern will be taken into account to the classifiers. Hidden Markov Model (HMM) based classifiers were proven to be suitable for facial expression recognition and have been widely used in this area. In Cohen, Garg, and Huang (2000), a multilevel HMM was proposed for automatic segmentation and recognition of human facial expression from video sequences. What kind of classifier to be used in the emotion recognition step depends on the face detection method and the extracted features.

CHAPTER IV

FRAMEWORK OF MULTI-SENSORY EMOTION RECOGNITION

In this chapter, the framework of multi-sensory emotion recognition system will be represented. As discussed in Chapter I, emotion recognition can be done via speech voice, speech content, facial expressions, as well as hand and body gestures. Currently, most of the approaches for emotion recognition mainly use one of these sources. In this dissertation, we tried to use the speech features and facial expression features to perform multi-sensory emotion recognition. The speech information can be the audio file recorded from the built-in microphone and the video streams recorded from the video sensor, such as the build-in Webcam of a laptop or other video camera. Recorded audio and video files can also be used as the input of the system.

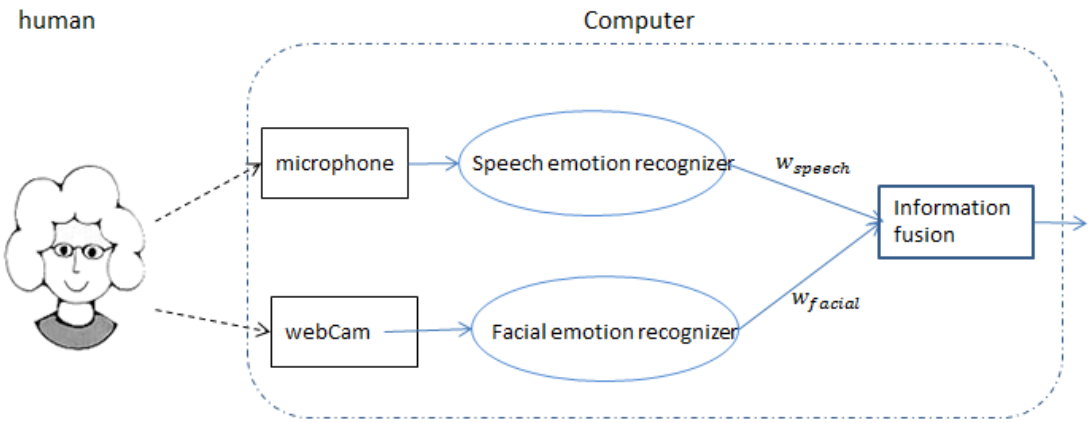

Figure 9. The framework of multi-sensory emotion recognition system.

There are three major parts of the system: the facial emotion recognizer, the speech emotion recognizer, and the information fusion. The core of this system is the information fusion part in which the emotion recognition results from the speech and facial expression will be integrated together with the Albert's communication model. After the resulting fusion, the final decision of the emotion will be given. From Figure 9,

we can see the general framework of the system design. The details of this work will be introduced separately in the subsection of this chapter.

The Implement of Speech Emotion Recognizer

As described in Chapter II, for the speech emotion recognizer, usually three steps are needed: data preprocessing, feature extraction and selection, and classification. Typical automatic emotion recognition from speech can be described in Figure 10 (Vogt, 2010). First, an emotional speech database is needed for the training and testing purposes. Suitable acoustic emotion units should be segmented from the instances. Then, the acoustic features should be fast extracted, and the most relevant features for emotion recognition should be selected. Finally, a fast and accurate classification algorithm is needed to classify the training data and build a base model. In the classification step, the emotion of test instances or real-time instances will be predicted according to the training model.

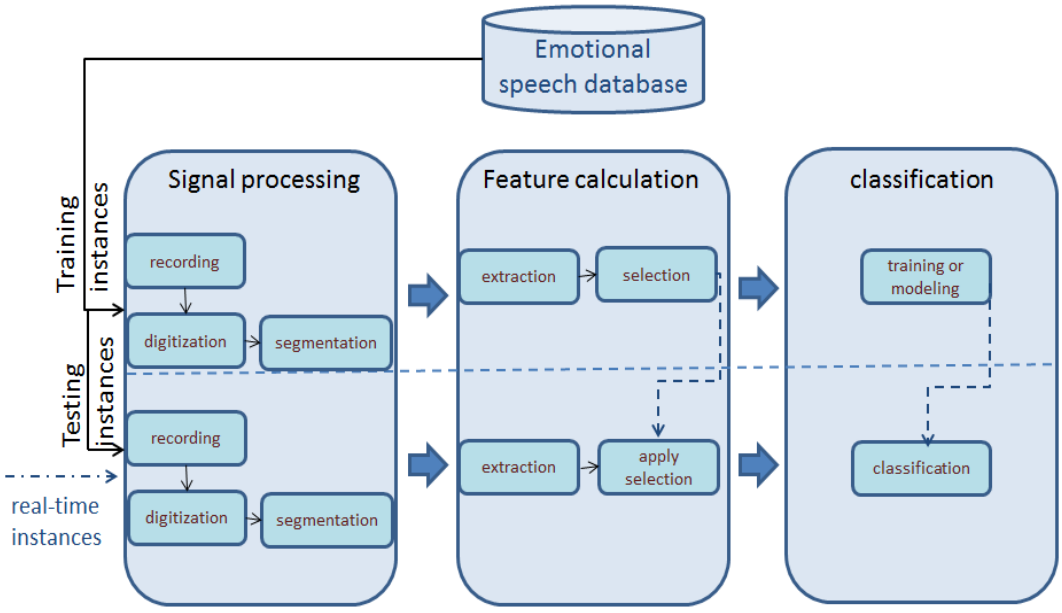

Figure 10. Typical steps for speech emotion recognition (Vogt, 2010).

Signal Processing

As introduced in the previous chapter, in signal processing step the analog speech data will be converted to digital data. Then, the digital data will be segmented into appropriate units. Though there are many types of the units, fixed size frames are widely used in many well-known available toolkits because of their efficiency. Figure 11 gives a better understanding of the signal processing.

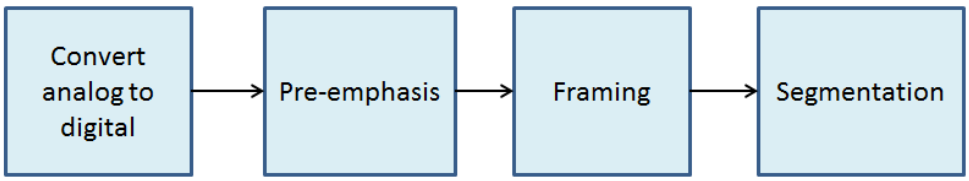

Figure 11. Speech signal processing.

The fundamental step is the analog to digital conversion. Currently, most audio recorders and players have the built-in facility to do this. However, advanced users can adjust sampling frequency and sampling resolution to improve the quality of the input speech data. The sampling frequency must satisfy the *Nyquist theorem* $f_s \geq 2f_c$, where f_c is the highest frequency contained in the signal, and f_s is the sampling rate. This theorem sets the lower bound for the sampling rate. While the sampling resolution is related to the precision of the signal sampled. For the simpleness of our design, we take the advantage of the built-in audio recorder of the laptop to record the speech.

Pre-Emphasis. After the analog signal is converted to the digital signal, the first thing to do is mean normalization also known as dc-offset elimination. The signal mean value μ_s should be found by the dc-offset elimination. The new signal after dc-offset elimination calculation is given by

$$s'[n] = s[n] - \mu_s \qquad (8)$$

in which $s'[n]$ is the modified signal after the dc-offset removal, and $s[n]$ is the original signal.

In pre-emphasis, the magnitude of the higher frequencies was increased according to that of the lower frequencies. The filter used here is the first order Finite Impulse Filter (FIR) as follows:

$$\mathrm{H(z)} = 1 - \mathrm{k}z^{-1} \tag{9}$$

where k∈[0.9, 1]. The first order differentiator output of the filter is signal $s'[n]$ which is simply the difference between two neighboring samples:

$$s'[n] = s[n] - ks[n-1] \tag{10}$$

Framing. Though speech signal changes continuously, to simplify and make the parameter estimation feasible, the speech is assumed to be stationary. Thus, the speech signal is divided into shorter segments with the assumption that locally the signal in each frame approximates a stationary signal. The most important parameters for the frame are length l_{fr}, overlap p_{fr}, and frame shifts$s_{fr} = l_{fr} - p_{fr}$. Usually, the frame length is around 20-25 ms which means there are 160-200 sample for f_s = 8000Hz. The overlap is used to avoid information loss.

Segmentation and windowing. Windowing function is used to select a frame and reduce the effect of spectral irregularities of framing. A suitable window function can smooth the edges of the frame in order to reduce the effect of discontinuities. The simplest window is the rectangular window. This kind of window has no change in the signal and is only used for selection. The following equation is the rectangular window equation.

$$w[n] = \begin{cases} 1 & if\ 0 \le n \le l_{fr} - 1 \\ 0 & other\ wise \end{cases} \tag{11}$$

The hamming window is preferred to use in the discontinuous case. The equation of Hamming window is

$$\mathrm{w[n]} = \begin{cases} 0.54 - 0.46cos\frac{2\pi n}{l_{fr-1}} & if \ \leq n \leq l_{fr-1} \\ 0 & otherwise \end{cases} \quad (12)$$

Feature Extraction and Selection

Two types of acoustic features will be used in this work including functionals and LLDs. Since the feature set is high dimensional, sometimes the feature selection algorithm is needed to select the most important features to reduce the calculation. The most famous feature selection algorithm is a greedy algorithm called the forward selection algorithm. It starts with an empty set and adds optimal features by some heuristic functions to this set iteratively.

In this dissertation, we extracted 384 features according to the INTERSPEECH2009 Emotion Challenge feature set (Schuller, Steidl, & Batliner, 2009) and 988 features according to the Emo-DB analysis as the feature sets for our speech emotion recognition. The SVM SMO and Naïve Bayes classifier are used for base model training and speech emotion recognition. Currently, the Naïve Bayes classifier gains many interests of researchers because of its high speed, good accuracy, and simplicity. It is a probabilistic classifier and works well in any domain. Suppose we have an n dimensional feature space, random variables Y and X_1, X_2, …, X_n; the feature vector components are x_1, x_2,…, x_n. The joint probability can be modeled

$$P(Y = y, X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \quad (13)$$

for any label y paired with x_1, x_2,…, x_n. It is possible with the Naïve Bayes assumption:

$$P(Y = y, X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(Y = y) \prod_{i=1}^{n} P(X = x_i | Y = y) \quad (14)$$

Naïve Bayes classifiers have many successful applications in Natural Language Processing, for example the text classification. The aim in using this classifier here is that it is fast and simple which may be good enough for the real time emotion recognition needs.

Since emotion recognition is a pattern recognition problem, the SVMs are the most used classifiers in this area. The SVMs are robust in solving high dimensional problem. Our problem is a multiclass classification problem. The SVM Sequential Minimal Optimization (SMO) classification method in Weka is used in this research. Compared to conventional SVM, the SMO algorithm is more efficient. The kernel used in SVMs in this research is the linear kernel.

Building the Base Model

In order to recognize emotion from speech, a base model should be built first. In this dissertation, since we desire to realize multi-sensory emotion recognition, we have not merely used an emotional speech database, but also used a multi-media emotional database (IEMOCAP) which includes audio and video information.

The Emo-DB database (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005), a benchmark in speech emotion recognition, is used to build a base model for the speech emotion classification. This database is free; anyone can download it from this website (Berlin Database of Emotional Speech, 2014). It contains around 535 speech files in wav format recorded by 10 German actors in 7 emotional states (happiness, anger, anxious, fearful, bored, disgust, and neutral). 10 German sentences used in everyday communication were used in this database, including 5 short sentences and 5 long sentences.

The IEMOCAP (Busso et al., 2008) was collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). It was recorded by 10 actors in dyadic sessions providing facial expression, hand movements, and spontaneous communication. The spoken language is English. It contains almost 12 hours of data. We can download from their website with permission. In our base model, we only used a small part of the audio file (more than 1000 sentences with 7 basic emotions).

After downloading one database, we keep all the files in a folder named corpus. Then we created 7 subfolders in corpus such as corpus/happiness, corpus/anger, and so on. Also, we save the speech files according to their names into the subfolders. Finally, each of the subfolders contains all the utterances for that particular emotion. This form of arrangement will be useful for us to perform the batch extraction and labeling of emotions during the feature extraction step.

In this research work, the Munich open Speech and Music Interpretation by Large Space Extraction (openSMILE) (Eyben, Wollmer, & Schuller, 2010) is employed to do speech signal processing, feature extraction, and real time input classification. OpenSMILE is a flexible feature extractor, especially for speech and music signal processing, and it can also be applied in many types of signal analysis. It was first presented at INTERSPEECH'09. It is written in C++ and can work on most platforms such as Windows, Linux, and MacOS. It can do general audio signal processing, speech related feature extraction, music-related feature extraction, and also contains many classifiers and other factors. It cannot only process large datasets in off-line batch mode, but also process the small data on-line with the PortAudio (PortAudio, 2014) library. In

this dissertation both on-line and off-line functionality are used to process the recorded information from databases and live data from the audio recorder.

Before extracting features the primary work is to make a proper configuration file. It is a type of file which is kept in a simple and powerful INI-style file format. In this configuration file, the user needs to specify the specific parameters such as the input information, output information, command line arguments, and the features to be extracted and so on. This file gives us the flexibility to choose the input data format, specify which features to be extracted, as well as the format of the output. Several output formats are supported, such as the ***arff*** format for Weka (Hall et al., 2009), ***htk*** format for the Hidden Markov Toolkit, the comma separated value text (***CSV***) format, and the LibSVM feature file format, etc.

In this research three configuration files were used and named *emo_IS09_new.conf*, *emobase_new.conf,* and *emobase_live4_new.conf*. The *emo_IS09_new.conf* extracts 384 features for static training, while the *emobase_new.conf* extracts 988 acoustic features. *emobase_live4_new.conf* is much more complicated, and it extracts features according the classification models and predicts the emotion of the speech lively via the input of PortAudio. In this dissertation, we test both models generated by *emo_IS09_new. conf* and *emobase_new. conf* and finally use the *emobase_new.conf* and *emobase_live4_new.conf* for on-line speech processing.

The SMILExtract binary command line utility is used for extracting features of a speech file. For building the base model we can use the Perl scripts, *build.pl*. To run it, Perl 5 needs to be installed. It can automatically extract features and build a classifier model for the features; however, the input must be available in uncompressed wave

format, 16 kHz sampling rate, mono or stereo, and saved in a corpus. All the speech data of one emotion class should be in a folder with the same class name. For example, all the instances of sadness should be in a folder named *sadness/*, like *sadness/sample_sadness1.wav*, and *sadness/sample_sadness2.wav*. After converting the format of the speech data, we can run as follows:

perl makemodel.pl path/to/the/corpus/ emobase_new.conf(emo_IS09_new.conf)

After running the command, the model is built, and there are several files in the folder of "build-model". Then, we can load this model, scale, classes, etc. to the live configuration file by creating a new cLibsvmliveSink section. Also, a new instance entry for the cLibsvmliveSink is needed.

The *makemodel. pl* is used to build the base model simply; however, the detail of the base model should be analyzed. Thus, we need to run another Perl script - *stddirectory_smileextract. pl* to extract only features of the corpus in batch mode and get the output in an *arff* file (such as *base.arff*). Then, another data mining software Weka (Hall et al., 2009) is needed to analyze the *arff* file to get a better base model.

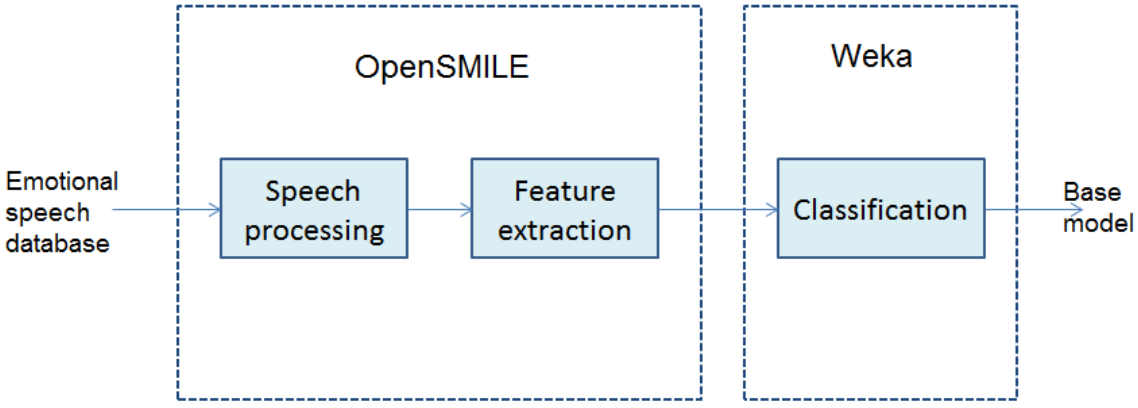

Figure 12. Building a base model.

The procedure of building a base model with OpenSMILE and Weka is given in Figure 12. We can also use Weka to do emotion recognition for off-line speech files. The detail will be discussed later.

The open source software Weka (Hall et al., 2009) under the GNU General Public License is used here to analyze the data for the base model. It is a collection of machine learning algorithms for the data mining task and contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It also supports the development of new machine learning algorithms.

The output feature file is used in Weka as the input and trained to an SVM model with the SMO algorithm and a Bayes model with the Naïve Bayes algorithm. 10 fold cross-validations were used for the model evaluation. SMO is used for solving the quadratic programming (QP) problem. It is widely used in training SVMs and is implemented by the popular LIBSVM tool (Chang & Lin, 2011).

Table 4

Confusion matrix of the SVM base model for Emo-DB (384 features)

Labeled	Recognized Emotion						
Emotion	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	112	0	0	1	14	1	0
Boredom	0	67	1	1	0	8	4
Disgust	1	2	38	2	1	0	2
Fear	4	0	3	57	3	1	1
Happiness	17	0	1	4	49	0	0
Neutral	0	6	1	1	1	69	0
Sadness	0	9	2	1	0	0	50

The overall recognition rate = 82.62%

Table 5

Confusion matrix of the SVM base model for Emo-DB (988 features)

Labeled	Recognized Emotion						
Emotion	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	118	0	1	1	7	1	0
Boredom	0	73	2	0	0	2	4
Disgust	2	1	39	2	0	2	0
Fear	3	0	1	60	3	1	1
Happiness	12	0	1	4	54	0	0
Neutral	1	3	0	0	0	74	0
Sadness	0	4	0	0	0	1	57

The overall recognition rate = 88.79%

Table 6

Confusion matrix of the Bayes base model for Emo-DB (384 features)

Labeled	Recognized Emotion						
Emotion	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	90	0	7	7	23	1	0
Boredom	0	49	9	2	0	12	9
Disgust	3	5	29	5	1	1	2
Fear	9	2	7	35	5	9	2
Happiness	19	0	1	6	41	4	0
Neutral	2	9	4	7	3	51	2
Sadness	0	8	1	2	0	1	50

The overall recognition rate = 64.49%

In the SVM model of the Emo-DB with the configuration of emo_IS09, 442 instances were classified correctly, and 93 instances were classified incorrectly. The recognition rate can achieve 82.62%. The confusion Matrix is given in Table 4. While

Table 7

Confusion matrix of the Bayes base model for Emo-DB (384 features)

Labeled	Recognized Emotion						
Emotion	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	20	1	2	1	104	0	0
Boredom	0	74	2	0	0	5	0
Disgust	0	12	23	1	9	1	0
Fear	2	5	1	29	29	3	0
Happiness	5	1	0	3	62	0	0
Neutral	0	37	0	2	10	29	0
Sadness	0	15	1	0	0	0	46

The overall recognition rate = 52.9%

Table 8

Emotion recognition rate with different base model for the Emo-DB

Emotion	*SMO Model* *(384 Features)*	*SMO Model* *(988 Features)*	*Bayes Model* *(384 Features)*	*Bayes Model* *(988 Features)*
Anger	0.875	0.922	0.703	0.156
Boredom	0.827	0.901	0.605	0.914
Disgust	0.826	0.848	0.630	0.500
Fear	0.826	0.870	0.507	0.420
Happiness	0.690	0.761	0.577	0.873
Neutral	0.885	0.949	0.654	0.372
Sadness	0.806	0.919	0.806	0.742
Overall Rec.Rate	0.826	0.886	0.645	0.529

Table 5 shows the confusion Matrix detail of the SVM model for the Emo-DB with 988 acoustic features. The recognition rate in this case can achieve 88.79%. However, in the Bayes model the recognition rate can only achieve 64.49% and 52.9% by using 384

acoustic features and 988 acoustic features, respectively. The details are shown in Table 6 and 7. From Figure 13, the comparison of different models, we can see that the SVM model with the SMO algorithm is much better than the Bayes model with the Naïve Bayes classification method. The model using 988 acoustic features performs better than the one using 384 features (See Table 8.). The same case is also suitable for the IEMOCAP database. We can see this in the later experiments.

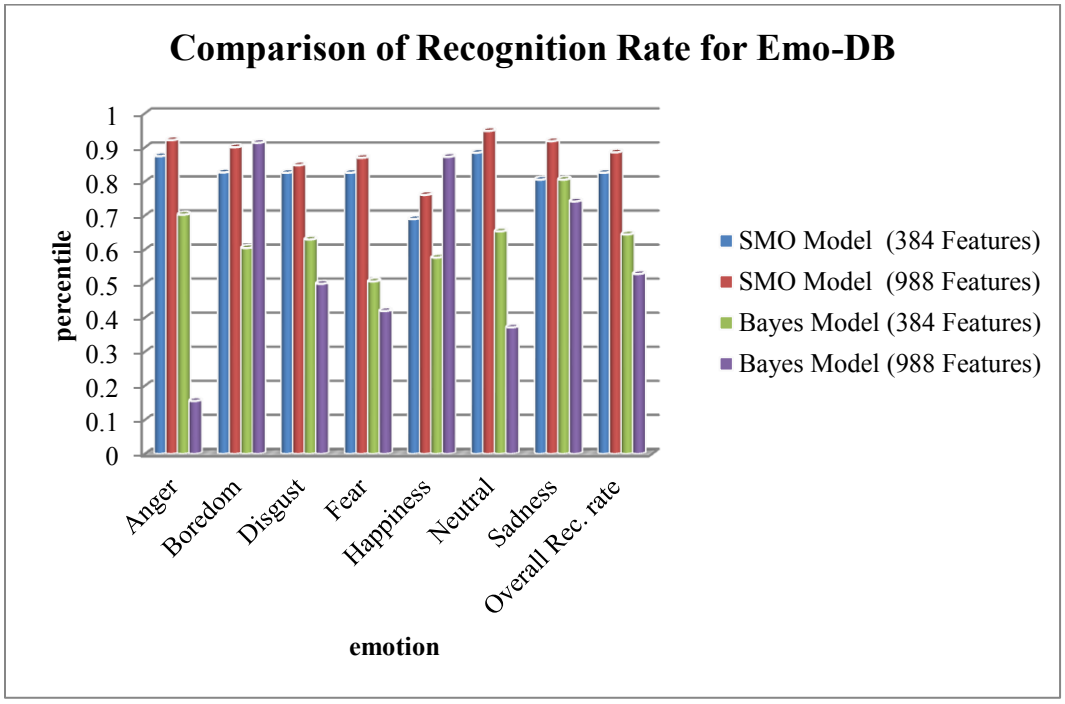

Figure 13. The accuracy comparison of different models for the Emo-DB.

For the multimedia database, if there are no separate audio files, we need to use the software *FFmpeg* to separate the audio files from the videos. FFmpeg is free software which provides libraries and algorithms for handling multimedia data. For the detail of this software, please refer to FFmpeg (2012). After the separation, the same thing will be done as in the Emo-DB base model building. However, the database IEMOCAP already has the audio files; we do not need to separate them ourselves.

Table 9

Confusion matrix of the SVM base model of the IEMOCAP (384 features)

Labeled	Recognized Emotion						
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	142	10	0	3	5	0	7
Disgust	9	34	12	14	1	3	18
Fear	2	13	43	4	15	7	21
Happiness	7	17	6	80	5	2	18
Neutral	6	2	5	5	143	42	2
Sadness	0	5	9	2	35	151	0
Surprise	13	22	24	21	10	8	94

The overall recognition rate = 62.63%

Table 10

Confusion matrix of the SVM base model of the IEMOCAP (988 features)

Labeled	Recognized Emotion						
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	145	14	0	1	3	0	4
Disgust	12	38	11	11	0	6	13
Fear	3	12	46	8	10	7	19
Happiness	2	18	6	88	5	2	14
Neutral	6	1	8	3	144	40	3
Sadness	0	7	16	2	30	143	4
Surprise	12	25	17	13	12	10	103

The overall recognition rate = 64.45%

In this experiment, 1097 instances were chosen from the IEMOCAP with different emotions (anger 209, disgust 115, fear 127, happiness 108, neutral 127, sadness 270, and surprise 141). Tables 9–12 show the detail of the confusion matrix of the

different base models for the IEMOCAP database. In the SVM model with 988 features, 707 instances were classified correctly with the highest recognition rate of 64.49%. From

Table 11

Confusion matrix of the Bayes base model of the IEMOCAP (384 features)

Labeled	Recognized Emotion						
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	136	6	1	10	0	0	14
Disgust	16	29	16	9	5	5	11
Fear	4	15	34	7	14	20	11
Happiness	16	28	18	50	1	1	21
Neutral	8	16	13	14	72	71	11
Sadness	1	6	11	1	22	157	4
Surprise	28	15	34	17	13	16	69

The overall accuracy = 49.86%

Table 12

Confusion matrix of the Bayes base model of the IEMOCAP (988 features)

Labeled	Recognized Emotion						
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	133	8	3	7	2	0	14
Disgust	14	32	14	13	3	7	8
Fear	4	23	25	7	10	25	11
Happiness	16	26	15	61	1	1	15
Neutral	10	27	6	15	54	82	11
Sadness	0	7	12	1	8	171	3
Surprise	28	16	29	21	6	31	61

The overall recognition rate = 48.95%

Table 13

Emotion recognition rate with different base model of the IEMOCAP

Emotion	*SMO Model* *(384 Features)*	*SMO Model* *(988 Features)*	*Bayes Model* *(384 Features)*	*Bayes Model* *(988 Features)*
Anger	0.850	0.868	0.814	0.352
Disgust	0.374	0.418	0.319	0.238
Fear	0.410	0.438	0.324	0.452
Happiness	0.593	0.652	0.370	0.263
Neutral	0.698	0.702	0.351	0.847
Sadness	0.748	0.708	0.777	0.318
Surprise	0.490	0.536	0.359	0.490
Overall Rec. Rate	0.626	0.644	0.499	0.352

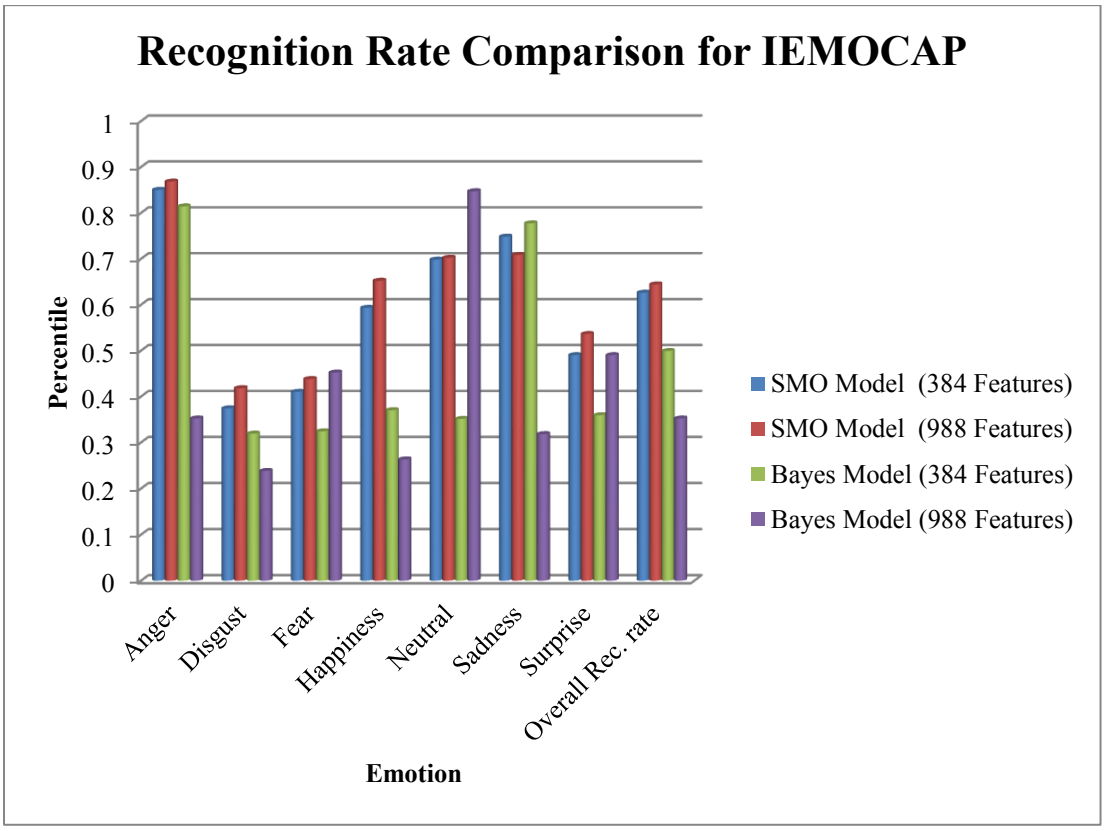

Figure 14. Recognition rate comparison of different models for the IEMOCAP.

Figure 14, we can have the same conclusion as for the Emo-DB experiments. The SVM model performs better than the Bayes model, while the recognition rate is higher when using 988 acoustic features. The details are shown in Table 13.

Above all, we can conclude that the base model using 988 features with the SMO algorithm has the best performance. The Bayes models have a much lower recognition rate than the SMO models. Thus, in the future we only use the SVM SMO algorithm with 988 acoustic features to build the base model for the speech emotion recognizer.

Facial Emotion Recognizer

In this section, the implementation of emotion recognition from facial expression is given. Two different approaches have been approached. The first one is mainly to process static images, while the second one is proposed for emotion recognition from videos or live stream from a webcam.

Facial Expression Recognition from Image

This approach can segment an input image by skin color and recognize facial emotion using the detection of facial feature and classification of emotion with simple curves and distance. The language used for this approach is the C# while MySQL database is used to store the training data. There are two important steps in this approach. It first detects and analyzes the facial region from an input image, and then recognizes the facial emotion of characteristic features in the region of interest. The details of the design for facial emotion recognition are shown in Figure 15.

In the face detection step, the system detects a face from the input image on the skin color and recognizes the region of the eyes and lip. It first extracts skin color pixels by initializing spatial filtering based on the result of lighting compensation. Then, the

system estimates the face position and the region of the eyes and lip by feature map. After obtaining the region of interest in the recognition step, the system will extract points of the feature map to apply Bezier Curve on the eyes and lip. The emotion recognition is performed by measuring the difference of Hausdorff distance with the Bezier Curve between input face image and the values stored in the database.

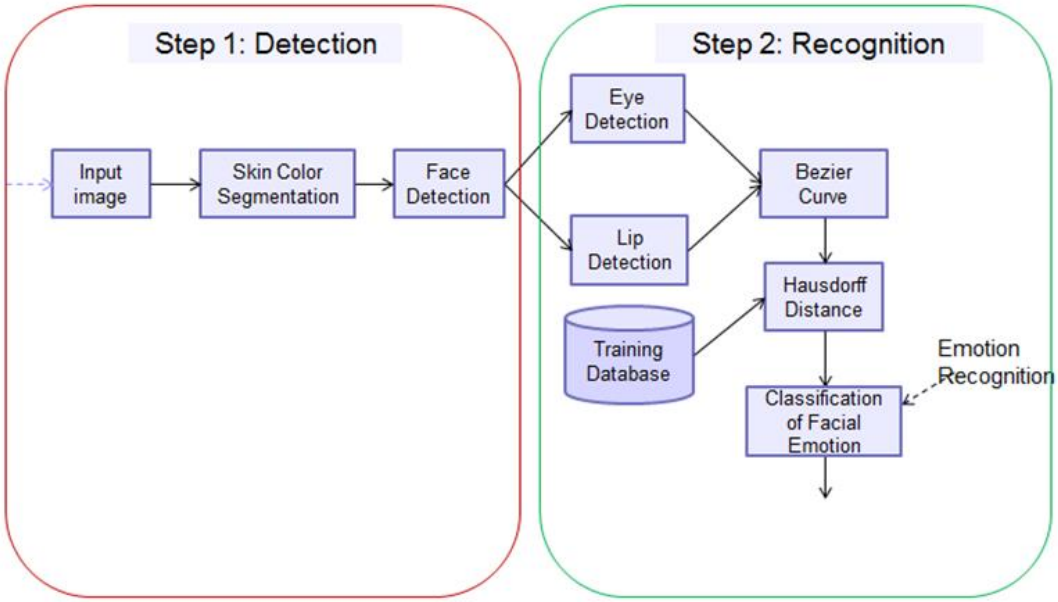

Figure 15. The first method of facial emotion recognition.

Input. The input for this approach is a static image.

Skin Color Segmentation. As we know the common features of human faces are eyes, lip, nose, eyebrows, forehead, cheeks, and hair. To detect emotion from face some features are required like eyes, lip and so on; while others are not, such as hair, forehead, etc. Thus, skin color segmentation is used to separate face and non-face elements of the entire image frame. Before segmentation the image needs to be contrasted to separate the bright and dark areas. After applying skin color segmentation, segments of adjacent pixels sharing similar visual characteristics are generated. Then, the system will find the

largest connected region. The probability of the largest connected region becoming a face will be checked. The rule is that if the largest connected regions' height and width are larger or equal than 50 pixels and the ration of height and width is between 1 and 2, then, this largest connected region may be a face.

Face Detection. Face detection will decide if there is a face or not in the image. First, the RGB image is converted to a binary image. In the conversion, we first calculate the average value of R, G, and B components of each pixel. If the average value is less than 110, the pixel will be replaced by a black pixel. Otherwise, the pixel is replaced by a white pixel. After these processes, the binary image is converted from the RGB image. Figure 16 shows the conversion of an RGB image to a binary image.

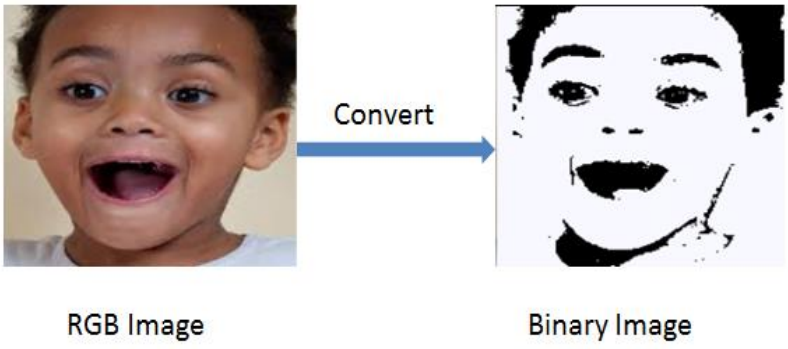

Figure 16. Convert an RGB image to a binary image.

In order to detect a face, we first scan from the middle of the binary image. The hair will come across first with a series of continuous black pixels. When it changes to white pixels, the forehead appears. This point will be marked and vertical scanning from the middle axis of the image on both sides, the left and the right, to find the maximum width of the face image. When scanning across the eyebrows, the black pixels are encountered, and the new width will be smaller than half of the previous maximum width.

At this point, the system will break, and the maximum width before encountering the eyebrows as the maximum width of the face is marked. We take 1.5 times of the maximum width as the height of the face. Thus, a rectangular of $1.5W * W$ (W is the maximum width of the face) will be cut from the point we have marked which only contains eyes, nose, lip, and eyebrows. Then the RGB image will be cut according to the binary image.

Eye Detection. The eye detector also needs to convert the RGB image to binary image and then scan image from W/4 to W-W/4 to find middle position M of two eyes. The highest white continuous pixel along the height between the ranges is the middle position of the eyes. By searching vertical, we can find the height of the starting position of two eyebrows. We search from W/8 to M to find the left eye and from M to W/8 to find the right eye. Some continuous black pixels are placed vertically from eyebrow to eye. To the left eye, the black pixels are placed in between M/2 to M/4, and to the right eye, the black pixels are placed between M+ (W-M) /4 to M+3* (W-M) /4. The height of the black pixel region is from H (H is the eyebrow starting height) to H/4.

To find the right side of the left eye and left side of the right eye, we search black pixels horizontally from the middle position to both sides between the upper position and the lower position of two eyes. The left side of the left eye is the starting width of the image, and the right side of the right eye is the ending width of the image. Finally, we map back and cut the left side, upper position, right side, and lower position of the two eyes of the RGB image.

Lip Detection. A lip box is used in lip detection step, and our assumption is lips must be in the lip box. The distance D_f between forehead and eyes should be determined

first. Then, the lower height of the eye H_l will be added to determine the upper height of the lip box. The starting of the lip box is from 1/4 position of the left eye and 3/4 position of the right eye. The lower height of the box is from the lower bond of the face image. The RGB image will be cut according to the lip box which includes the lip and some part of the nose.

Bezier Curve Application on lips and eyes. Bezier Curve can be used in many applications such as approximation, interpolation, curve fitting, object representation, and so on. Using this algorithm a new cure can be formed by finding points in the middle of 2 nearby points and repeated until there is no more iteration. The Cubic Bezier Curve is used in this system. Figure 17 shows the cubic Bezier curve of the four according to the four points P_1, P_2, P_3 and P_4. The formula is

$$B(t) = (1-t)^3 P_1 + 3(1-t)^2 t P_2 + 3(1-t) t^2 P_3 + t^3 P_4, \quad t \in [0,1] \qquad (15)$$

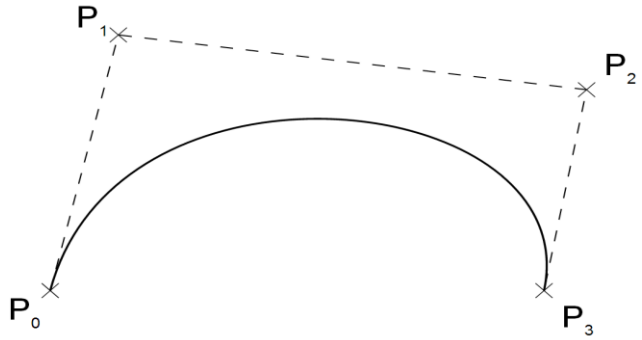

Figure 17. Cubic Bezier Curve.

In lip detection, we get the lip box contains lip and part of the nose. We will need to do some conversion to get rid of the noise around the lip. We will convert the skin pixel to white pixel and the other pixel to black. The pixels which are close to skin pixels are also converted to white pixels. In the RGB image, if the difference of two pixels RGB values is no bigger than a particular value, they are considered as similar pixels. A

histogram is used for finding the difference between the lowest and the highest RGB value. If the difference is less than 70, 7 is used for finding similar pixels; if the difference is larger or equal 70, 10 will be used to find the similar pixels. Thus, the particular value depends on the quality of the image. Usually, a high quality image uses 7 to find the similar pixels; otherwise 10 is used.

In the binary image of the lip box, lip is the largest thing except skin.Thus, big connected is applied to find the largest black region as the lip. After this, Bezier Cure is applied to the binary lip. First, we need to find the start and end pixel of the lip in the horizontal direction. Then two tangents are drawn on upper lip from the starting and ending pixel, and also two points are found on the tangent which is not the part of the lip. With the start and end points and two other points, we can draw the Bezier curve for the upper lip. A similar thing is done to the lower lip to form the lower lip curve.

Before applying Bezier Curve on the eyes, the eyebrows should be removed from the eye image. In this case, the first continuous black pixels are removed from the box. Then, the Cubic Bezier Curve will be applied to eyes as to the lips. The basic steps for applying Bezier Curve to the lip and eyes can be seen in Figure 18.

Figure 18. Bezier curve application for eye and lip.

Training and Recognizing Facial Emotion with Hausdorff Distance. There are two tables in the training database. One is used to store personal information and the index of the basic emotions, and another one is used to save control points for the Bezier

curves of lip, left eye and right eye, as well as the height and width of the lip, left eye, and right eye.

To recognize the emotion from an image, we have to find the Bezier Curve of the lip, left eye, and right eye. Then, we convert each width of the Bezier Curve to 100 and convert the height according to its width. Then, we apply the Hausdorff distance to compare the shape metric between them. The distance $d_H(p, q)$ between two curves $\mathrm{p(s)}, \mathrm{s} \in [\mathrm{a}, \mathrm{b}]$ and $\mathrm{q(t)}, \mathrm{t} \in [\mathrm{c}, \mathrm{d}]$is given in Equation (14) (Kim & Ahn, 2007).

$$d_H = \max\left\{ \underset{s \in [a,b]}{max} \underset{t \in [c,d]}{min} |p(s) - q(t)|, \underset{t \in [c,d]}{max} \underset{s \in [a,b]}{min} \left|p(s) - q(t)\right| \right\}. \quad (16)$$

The nearest matching pattern is picked, and the corresponding emotion is given. If the input does not match any one in the database, then the average height for each emotion in the database is calculated, and the decision is generated on the basis of average height. We created a dataset of feature points, height and width for the lip, left eye, and right eye by experiments. Then, from this dataset we find the value range of different feature points extracted from different facial emotions. According to this range, we can decide the facial emotion of the input image.

This approach performs well when the inputs are static images. However, to convert the process to recognize facial emotion with the video frames on-line or off-line is much complicated. Furthermore, it performs better when using small data sets, but time consuming when the training database is large. After doing lots of research, we found that the Open Source Computer Vision Library (OpenCV) can support image processing of videos on-line or off-line perfectly. Thus, we proposed the second approach for facial emotion recognition with OpenCV library and implemented with C++ using Microsoft Visual Studio 2010.

Facial Emotion Recognition from Video and Online Camera

Before introducing this method, let us see some detail of OpenCV first. OpenCV has been released under a BSD license which is free for both academic and commercial use. It can run on multiple platforms such as Windows, Linux, Mac OS, iOS, and Android. The library is written in optimized C/C++ and has more than 2500 optimized algorithms, including classic and state-of-the-art computer vision and machine learning algorithms. With these algorithms, OpenCV can be used to detect and recognize faces, identify objects, classify human actions in videos, track movements in cameras, track moving objects, etc. It is used worldwide in companies, the academy, and even governments.

In this approach of facial expression recognition, the Active Shape Model (ASM) based tracker, *asmlibrary* which was developed by Wei (2009) under OpenCV, was used and modified. The procedure is as follows. First, the modified ASM face tracker is used to track the landmarks. Then, the high level features are computed according to the landmarks, and seven regions are gotten as the high level features. After that, a histogram using Principal Components Analysis (PCA) coefficients of those high level features is created. Finally, the results of the seven emotion classes are generated with a distance based classifier.

Facial Landmark Tracking. As we introduced in Chapter III, ASM is one of the state-of-the-art approaches of facial landmark tracking (Cootes et al., 1995). To train the ASM annotated face images were used. A global face detector, Viola Jones face detector is used to locate the face. Searching of landmarks will start from the mean shape aligned to the position and size of the face. The following two processes are repeated until

convergence: (a) adjust the locations of shape points by template matching of the image texture around each landmark and then propose a new shape; (b) conform this new shape to a global shape model based on PCA. The single template matching is unreliable, and the shape model improves the results of the weak template matches by forming a stronger overall classifier. The entire search is repeated at each level in an image pyramid, from coarse to fine resolution using a multi-resolution approach. When tracking, the model is initiated from the shape found on the previous frame instead of using the mean shape. The mean shape is calculated in the initialization step.

It is well known that ASM has better performance when the model is trained person-specific. However, in order to use it for different persons, a generic model is needed. This generic model was built on some images of the Cohn-Kanade Expression Database (Kanade, Cohn, & Tian, 2000). The images were annotated manually during the building process. Also, a generic emotional model with 7 emotions was trained from the researcher's emotional face images. Nevertheless, the new person specific model can be added by training facial images of the person.

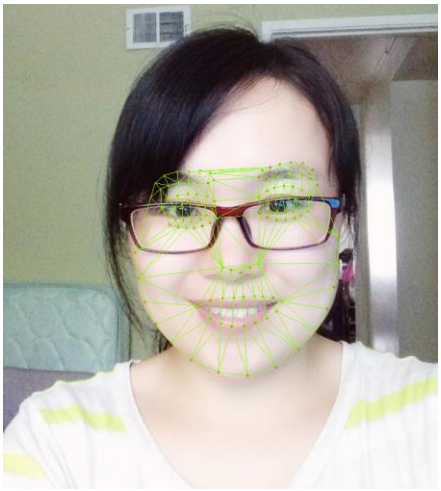

Figure 19. 116 Facial landmarks in shape.

Feature Extraction. There are 116 facial points used as the landmarks during the tracking step. A sample of the landmarks and shape is shown in Figure 19. The tracker works fine on most parts of the face such as eyebrows, eyes, nose, but is not very robust for detecting the exact movement of lips. The reason is that the lips are the most non-rigid part of the human being's face. When using the ASM tracker holistically, small variations in the lips may be discarded during the constraining of PCA. In addition, the intensity difference on the boundaries of the lips is not as obvious as other parts, and this can be seen in Figure 19. Thus, using the positions of the landmarks directly is not a good idea. To overcome the shortcoming of using the landmark positions, we compute seven significant features of the face with the landmark positions as the high level features. Figure 20 shows the comparison of 116 landmarks and the seven high level regions calculated from the landmarks. The video sample of Figure 20 is from Martin, Kotsia, Macq, and Pitas (2006).

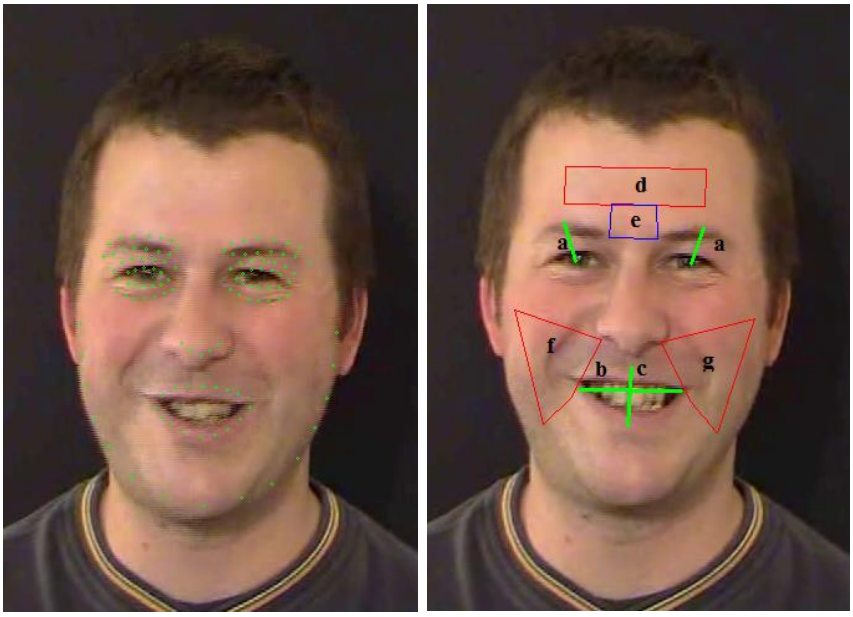

Figure 20. Facial landmarks (left) and the regions (right) for feature extraction (Martin et al., 2006).

The seven high-level features are as follows:

(a) Average distance from the eye middle of the eyebrow middle

(b) Lip width

(c) Lip height

(d) The vertical edge activity over the forehead region

(e) The horizontal edge activity over the lower forehead region

(f) Sum of horizontal and vertical edge activity over the right cheek

(g) Sum of horizontal and vertical edge activity over the left cheek

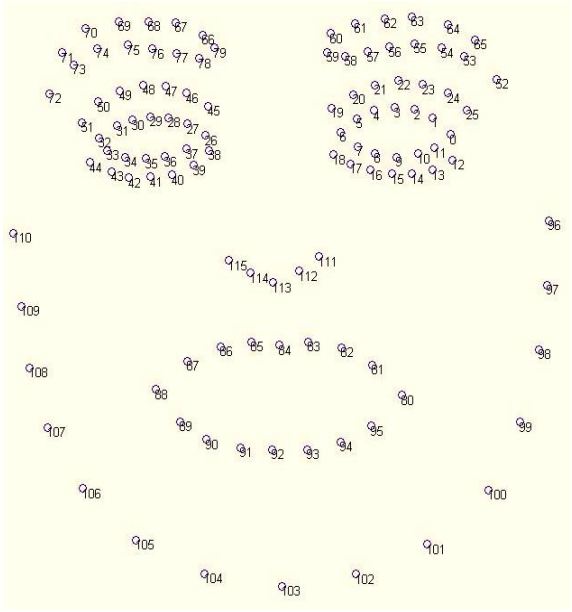

Figure 21. The facial landmark scheme.

Features a, b and c are computed using the Euclidean distance between the related landmarks. The order of the landmarks on a face is fixed, and this can be seen from Figure 21. To feature a), we need first to find the position of the middle of the right eyebrow (x_{rbr}, y_{rbr}), right eye (x_{reye}, y_{reye}), left eyebrow (x_{lbr}, y_{lbr}), as well as the left eye (x_{leye}, y_{leye}), and then calculate the distance between (x_{rbr}, y_{rbr}) and (x_{reye}, y_{reye}),

as well as (x_{lbr}, y_{lbr}) and (x_{leye}, y_{leye}). Points 55 and 63 were used to get the right eyebrow middle position (x_{rm}, y_{rm}). Points 69 and 75 were used to calculate the middle position of the left middle position (x_{lm}, y_{lm}). The middle position of the right eye is the middle of point 3 and point 9, and the middle position of the left eye is the middle the middle of point 29 and 35. Then, use the following equation to calculate the Euclidean distance.

$$D_{x,y} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{17}$$

To calculate the remaining five features the smooth function of OpenCV is used to blur the image with a Gaussian Kernel. Next, a Sobel kernel filter is used on horizontal and vertical axes. The vertical edge activity image for the neutral emotion is shown in Figure 22. Then, the average absolute pixel value is computed in each region. For instance, the average pixel value residing in the vertical edge activity image is represented by the g^{th} feature.

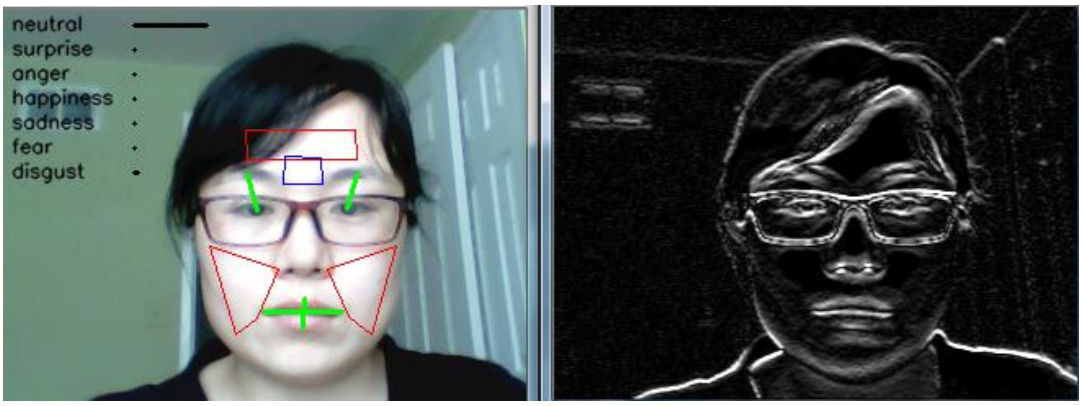

Figure 22. The neutral emotion (on the left) and its corresponding vertical edge activity image (on the right).

Smoothing is commonly used in image processing to eliminate noise. It is also called blurring. To perform a smoothing operation, a filter needs to be applied to the image. There are various filters can be used such as linear filters, the Normalized Box

Filter, the Gaussian Filter, the Median Filter, and the Bilateral Filter. Among these filters, Gaussian filter is the most useful one. It can be done by convolving each point in the input array with a Gaussian Kernel and then summing them all to produce the output array. The 2D Gaussian is used here in OpenCV, and it can be represented as:

$$G_0(x,y) = Ae^{\frac{-(x-\mu_x)^2}{2\sigma_x^2}+\frac{-(y-\mu_y)^2}{2\sigma_y^2}} \tag{18}$$

where σ_x is the Gaussian Kernel standard deviation in x direction and σ_y is the Gaussian kernel standard deviation in the y direction. The smoothing function is: *cvSmooth(src, dst, CV_GAUSSIAN, n, n)*. The parameters *src* and *dst* are the input image and the output image, respectively; CV_GAUSSIAN is the Gaussian Kernel like in the equation (18); while n is the parameters for the blurring which means convolving image with $n \times n$ Gaussian.

The discrete differentiation operator used here is the Sobel operator (Sobel operator). It computes an approximation of image intensity and combines Gaussian smoothing and differentiation. It also calculates the horizontal changes and vertical changes. Let us take a kernel size as 3 for example. Assume the input image is I, G_x and G_y are two images which at each point contain the horizontal and vertical derivative approximations, respectively. The calculations are as follows:

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I \tag{19}$$

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * I \tag{20}$$

Where $*$ here denotes the 2-dimensional convolution operation. By combining both G_x and G_y, the approximation of the gradient at each point of the image can be calculated.

$$G = \sqrt{G_x^2 + G_y^2} \quad (21)$$

Sometimes the simpler equation $G = |G_x| + |G_y|$ is used instead of Equation 21. The application of Sobel Operator can help to generate the detected edges on a darker background in the output image. The Sobel operators combine with the Gaussian smoothing, and differentiation in OpenCV can improve the result more robust to the noise.

In the setup step, the first m frames of the test subject are used to initialize in order to get the mean shape of the test subject's face. m is the threshold for initialization. The average values of the features are computed, and the features in the remaining frames are normalized. For the best recognition result, neutral expression is recommended for the initialization step.

Emotion Classification. The facial expression varies with human beings and the conditions of the surroundings such as light. The training setup can be easily configured for a specific individual in a specific environment. A menu was done on the interface; the test subject can choose what to perform next. For instance, one can add a new user and then train a new model to this user. Also, the test subject can choose to add new emotional model to the existing user, or he/she can use the existing user and existing emotional model to do facial emotion recognition. If one wants to build a fresh model when the training starts, the test person should wait in the neutral states and then repeats each state for specific times (such as 6). The interface records the feature vectors for these states, which are a total of 42 different vectors belonging to seven classes.

When doing the test, the test person should go with a neutral expression for the normalization and adaptation to the environment. The feature vector will be computed for

each frame, and then its average distance to the training feature vectors are computed for each class. Assume $d_i, i = 1, \dots, 7$ are the average distance computed for each class. Here, $D_i = e^{-d_i}$ is used as a measure to represent the distance dissimilarity. The values of D_i are normalized as their sum is one, and they can be seen as likelihoods.

Reference to New Subjects. Training an ASM model for a new person needs to collect 30 images of this person. The generic model is practiced for the automatic landmarking of the faces instead of annotating them manually. Then, the locations of the landmarks are fine-tuned, and the person specific model is trained from them. The training of the classifier can also be done easily for a specific person under the current surroundings. The process can be extended to new users. When adding a new subject, the subject can choose to add a new user from the menu of the interface. Then, the interface will take this subject's photo and annotate the landmarks on the photos automatically. Also, we can adapt the landmarks manually. The information about the annotation will be saved in an Additive Manufacturing File Format (amf) named *username.amf*. The test subject needs to add new expression class to the new user as shown in Figure 23. Then it will ask the subject to decide how many emotions (such as 3) are in the new class and how many samples (such as 5) are there for each emotion and what kind of emotions. After all the samples are taken, the new emotional model for the new user is set up. Then, run the program again and chose to use the new user and new expression class for facial emotion recognition. The results of a toy example are shown in Figure 24.

Results. The results of this facial emotion recognition design are shown in Figure 25. The likelihoods of the related emotional states are drawn on the user interface and also saved in a specific field for other use, such as information fusion or result

comparison. A Core2 Duo CPU at 2.53 GHz laptop computer with 4 GB RAMS and 64-bit Windows 7 operating system was used. It can process up to 30 frames per second with 640 ×480 pixels resolution. The built-in laptop webcam is used, and we can also use the other camera device to perform the training and facial emotion recognition.

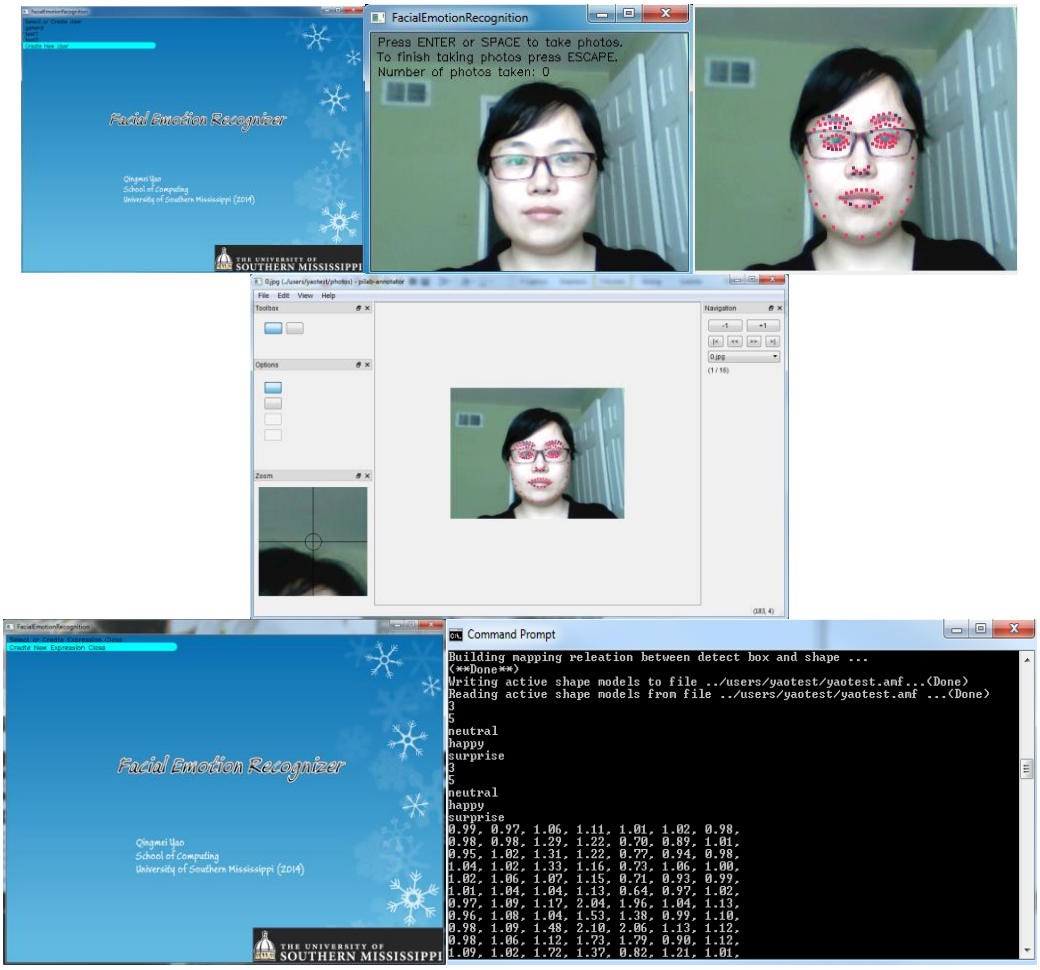

Figure 23. The procedures for adding a new user and adding new emotional classes.

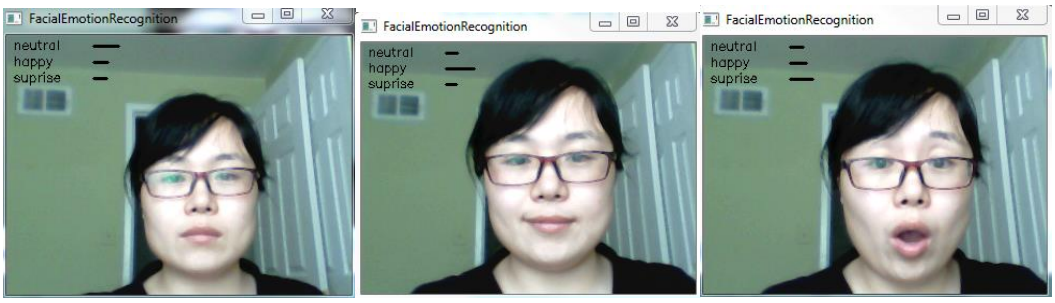

Figure 24. A toy example for 3 emotion recognition.

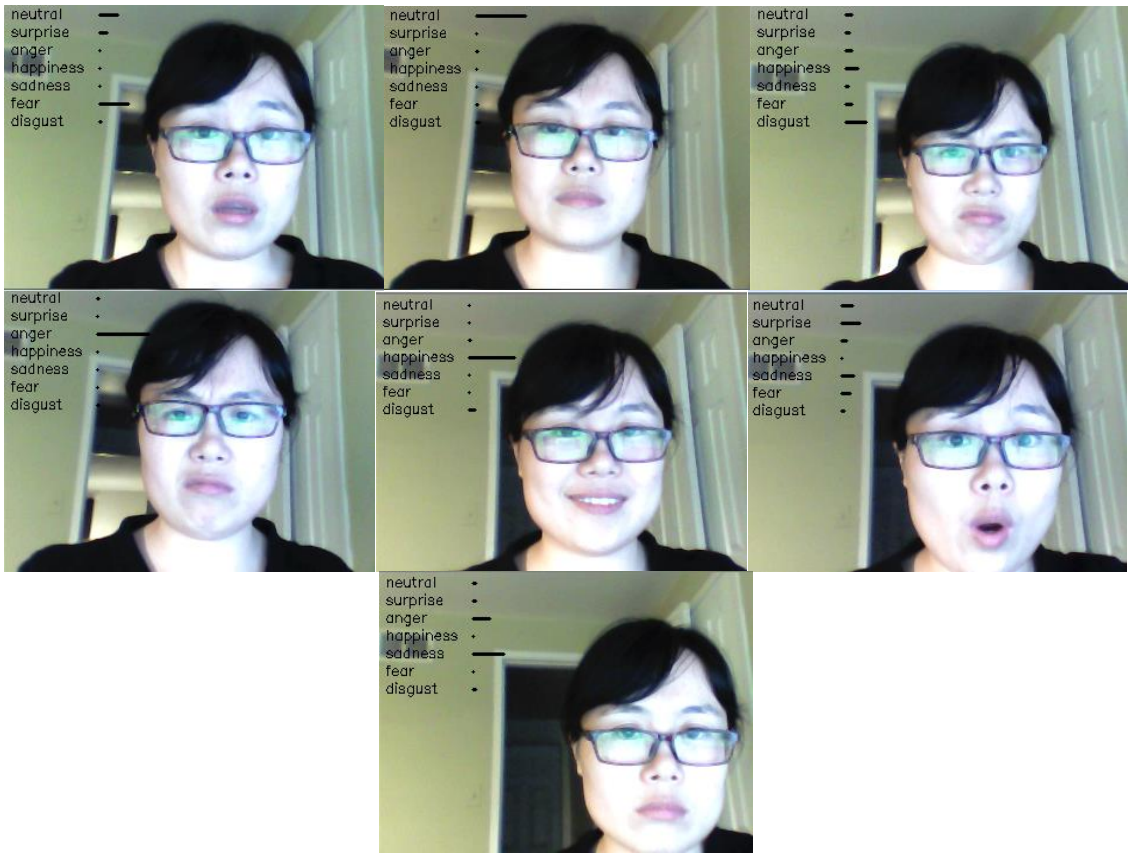

Figure 25. Facial emotion recognition results in recognizing 7 emotions (fear, neutral, disgust, anger, happiness, surprise, and sadness).

Results show that this facial emotion recognizer can also handle some of the partial facial emotions as shown in Figure 26.

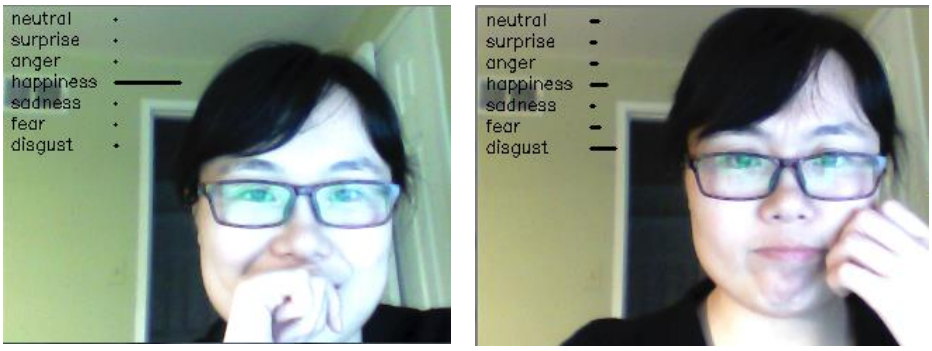

Figure 26. Emotion recognition results for partial face.

The above figures show the facial emotion recognition result of each frame. In order to get the emotion of a video clip, which means for a period of the video, we used

the majority method. There is a counter for each emotion; then we count the emotion of each frame. The emotion with the largest number will be chosen to represent the emotion of the video clip.

Information Fusion

In the process of emotion recognition, information fusion refers to combining and integrating all incoming information into one representation of the emotion expressed by the user. There are two problems to solve in this step: when to integrate the information and how to integrate the information (Corradini, Mehta, Bernsen, Martin, & Abrilian, 2003). Generally, there are two major information fusion methods: feature level fusion and decision level fusion.

Feature level fusion is done on the feature sets extracted from the speech and facial expression, and the process is shown in Figure 27. However, the feature level fusion cannot be normalized well when the information is different in the temporal characteristics. One requirement for the extracted features is that they should be synchronous and compatible. Another problem to be figured out for feature level fusion is the high dimensional data which can result in heavy computing and time consuming.

Decision level information fusion is based on the assumption that different modules are independent from each other. In this kind of method, each module (in our case is speech emotion recognition or facial expression recognition) is classified separately, and the output of each module is integrated to get the global decision of the expressed emotion (see Figure 27). Currently, the decision level information fusion method is widely used for multimodal emotion recognition.

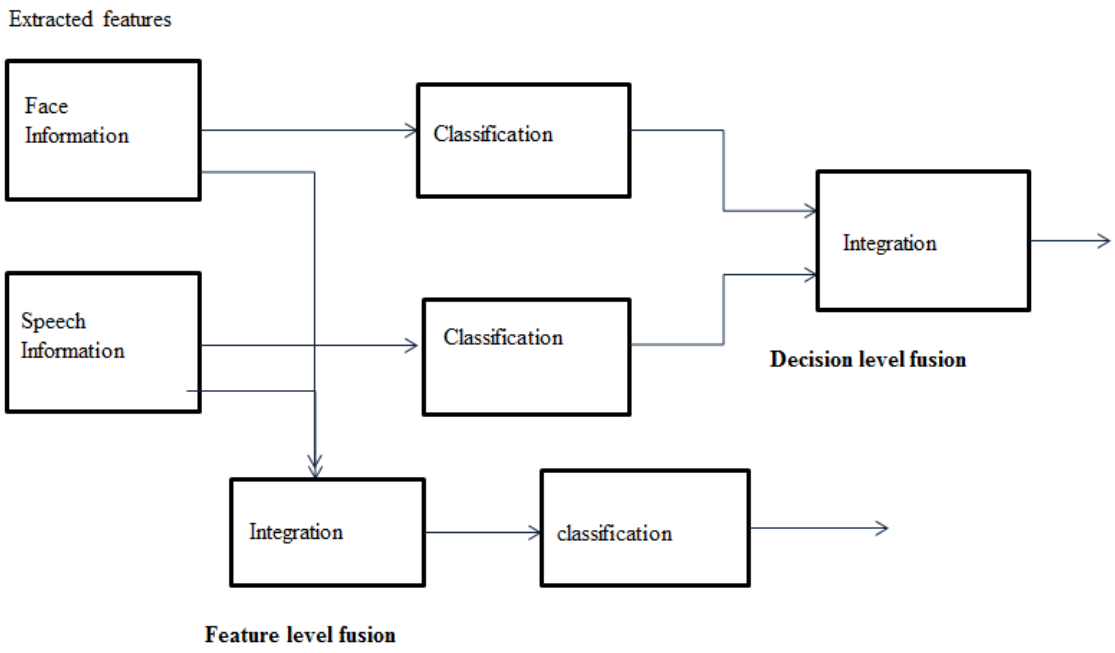

Figure 27. Overview of two information fusion methods.

There are many reasons to use decision level information fusion instead of feature level information fusion (Wu, Oviatt, & Cohen, 1999). The feature level fusion needs to use large multimodal dataset because the feature sets are in a high dimensional data space. In the decision level fusion, multi modules can be processed asynchronously. Besides, it provides more flexibility. Thus, we can use different classifiers on different data sets and integrate them without retraining. By using the decision level fusion, the recognizer can be used for single module emotion recognition. Moreover, it is more flexible and extendable using the decision level information fusion. Therefore, in this dissertation, we chose the decision level information fusion in the multi-sensory emotion recognition.

When doing the speech emotion recognition, the results of the recognized emotion in each turn were saved in a text file with a name such as speech_v1.txt. The records in the file include the number of the recognized emotion, the starting time of each turn, the duration of each turn, and the probabilities of the seven emotions during this period. The

results of the facial emotion recognition were also saved in a text file. In this file there includes the frame number and the probabilities (likelihood) for the seven emotions. The frame number starts after the initialization threshold.

Fusion Algorithm

Load and save the speech emotion recognition result

Load and save the facial expression recognition result

Calculate the start point for the multi-sensory emotion recognition T_{start}

Information fusion from T_{start} to the end of the video T_{end}

a. Calculate the durations
b. Use the emotion of the speech turn in that duration as the speech emotion E_{speech} and its probability P_{speech}
c. Calculate the likelihood $L_i = \frac{C_i}{\sum_{i=0}^{n} C_i}$ $(n = 0, 1, 2, \ldots 6)$ of the emotions from the video frames, where C_i is the counter to count the number of those seven emotions during this duration, the largest likelihood L_i will be set as the probability P_{facial} and the relevant emotion as the recognized facial emotion of this period E_{facial}.
d. Assign weights to the speech emotion as $w_s * P_{speech}$ and facial emotion as $w_f * P_{facial}$ and then compare the two; the larger one will be set as recognized emotion and probability

The first duration is from T_{start} to the start of the next turn no matter when the first turn starts.

Other durations are from the start of the turn to the start of the next turn.

The last duration is from the start of the last turn to the end of the video T_{end}

The difficulty for the information fusion is that the formats for speech emotion recognition and facial expression recognition are different. The speech emotion

recognition is for each turn, and there are blanks when the turn stops or before the starting. But the results in facial emotion recognition are frame by frame. In order to fuse the results together, we have to know the frame rate of the video and then calculate the emotion.

The information fusion can be done as in the *Fusion Algorithm*. The details are as follows:

(a) Load the emotion recognition results from the speech file, save the recognized result, starting time, and duration in one-dimensional arrays, respectively. Save the probabilities of seven emotions for every turn in a two-dimensional array

(b). Load the results from the facial emotion recognition part, save the frame number in a one-dimensional array, and save the probabilities of the seven emotions in a two-dimensional array.

(c). Calculate starting time of the emotion recognition according to the starting number of the video frame and frame rate. Then compare the starting time with the turns. From Figure 28, we can see that the situation of the turns of the speech and the video frames are complicated. To simply the fusion, we set the start time of the facial expression recognition as the start time of the whole multi-sensory emotion recognition. Also, set the time of the last frame of the video as the final end.

(d) After getting the start point, we will calculate the total frames F_i from one turn start to the start of the next ($duration_i$) like in Figure 29. The emotion from the speech is the emotion of *turn i,* and the emotion from the facial expression will be calculated from the video frames F_i. We will compare the probabilities of the seven emotions for each frame and find the emotion with the largest probability. There is a counter for each emotion;

when an emotion with the largest probability is found, one will be added to the counter of this emotion. After all the frames in F_i are processed, the emotion with the largest number in the counter will represent the facial emotion of the video in the $duration_i$.

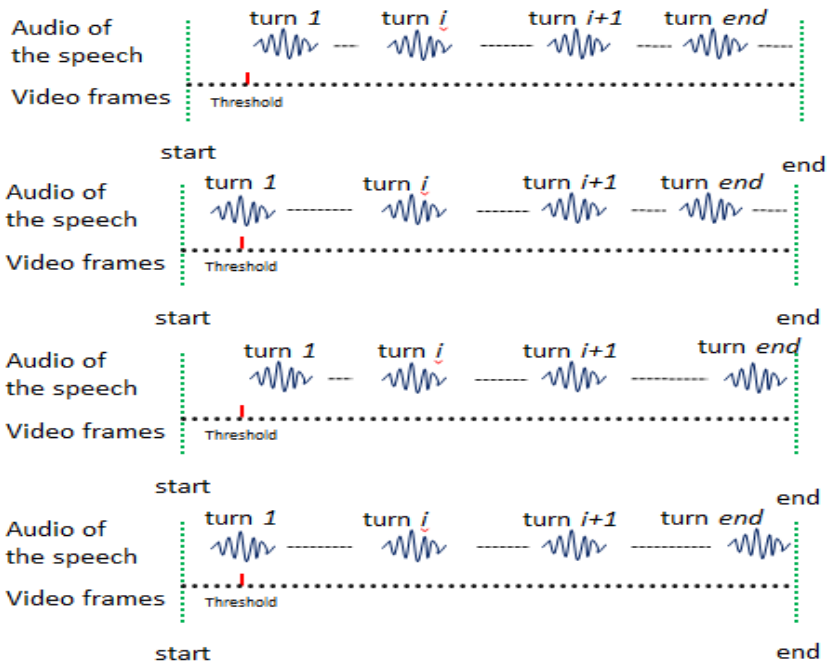

Figure 28. Some of the situations for the Audios and videos.

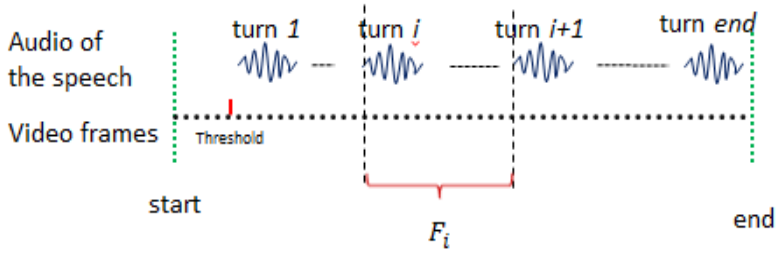

Figure 29. Calculating the video frames.

(e) Finally, we come to the fusion. In this step, we add two coefficients w_f and w_s to the results from speech and facial expression, respectively. There are two approaches for the information fusion. The first one integrated the results of the speech emotion recognition and facial expression recognition directly which means $w_f = 1$ and $w_s = 1$ in the fusion.

While the second approach adds weight to the two results according to Albert's communication model (Mehrabian, 1971). Albert's research concludes that during the communication between human beings, 55% of emotional feelings and attitudes is represented in facial expression; 38% is shown in the paralinguistic (the acoustic feature of speech); and only 7% is related to the words that are spoken. Thus, in the fusion algorithm w_f = 0.55 and w_s = 0.38. The 0.07 was omitted since we do not concern the content of the speech.

Summary

In this chapter, the approaches for speech emotion recognition, facial expression recognition, and information fusion have been introduced in detail. The framework of the multi-sensory emotion recognition is shown in Figure 9. For the speech emotion recognition there are three steps: speech processing, feature extraction and selection, as well as the classification. All these works can be done by the software OpenSMILE with our specific configuration files. In order to get a more stable base model for the speech emotion recognition, Weka was use to analyze the extracted features, accuracy, and other details of the results from two different emotional speech databases, Emo-DB and IEMOCAP. The results show that using 988 features with the SMO classification method can achieve better performance than using another feature set or classification method. There are two achievements in facial emotion recognition. The first approach is used to recognize facial expression from static images by the techniques of skin color segmentation and Bezier Curve. While the second one is based on the ASM face tracking and is perfect for recognizing facial emotion from videos or live streams of a Webcam. No matter which achievement, there are some common steps, and they are face detection,

feature extraction, and classification. Finally, as shown in Fig 9, the information fusion was done with and without the adding weight. It is the first time to add weight in the information fusion. With the design of our multi-sensory emotion recognition, the emotion recognition accuracy has been improved 18.27% compared to the speech emotion recognition and 5.66% compared to the facial emotion recognition. The details of our experiments and results are shown in Chapter V.

CHAPTER V

SYSTEM TESTING

So far, the framework of the multi-sensory emotion recognition has been introduced, including the speech emotion recognizer, facial expression recognizer, and information fusion. In this chapter, many experiments have been done based on the proposed work, and the results were examined in detail.

Choosing Base Model for Speech Emotion Recognizer

In the speech emotion recognition approach, we have already built many different models based on different databases, feature sets, as well as the classification algorithm. According to the conclusion, the Emo-DB and the IEMOCAP models based on the SVM SMO algorithm with 988 acoustic features were used. The tests in this section were done with Weka and OpenSMILE; the process is presented in Figure 30. As we know the language in the Emo-DB is German and in the IEMOCAP is English. The following tests will also show how the language affects the accuracy of the speech emotion recognition.

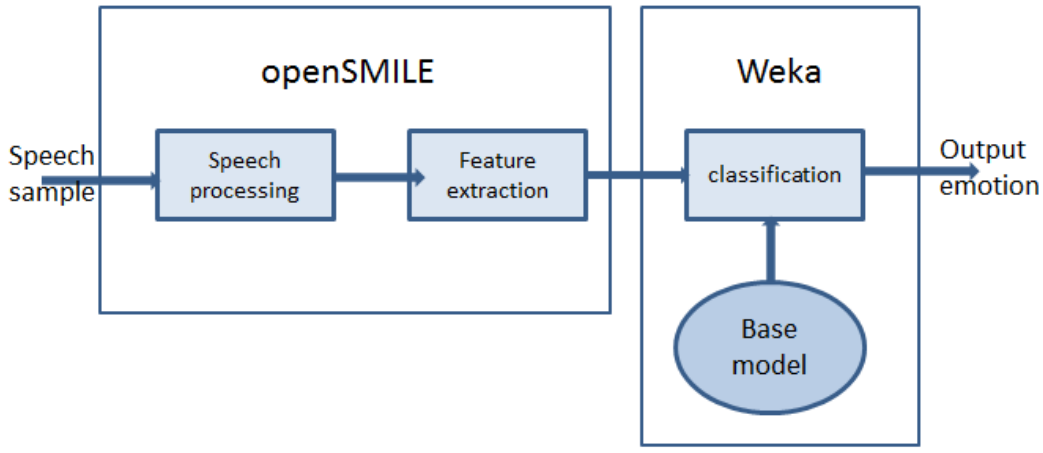

Figure 30. Speech Emotion Recognition.

Table 14

Confusion matrix of the SVM Emo-DB base model with 491 instances (384 features)

Labeled	Recognized Emotion						
Emotion	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	109	1	3	6	1	0	0
Boredom	0	66	1	0	0	4	3
Disgust	1	1	35	3	0	1	0
Fear	4	0	1	55	2	0	1
Happiness	11	0	1	4	48	1	0
Neutral	0	5	1	0	1	64	0
Sadness	0	7	0	0	0	1	49

Overall recognition rate = 86.76%

Table 15

Confusion matrix for the 46 instances of the Emo-DB (988 features)

Labeled	Recognized Emotion						
Emotion	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	7	0	0	0	1	0	0
Boredom	0	7	0	0	0	0	0
Disgust	0	0	4	0	0	1	0
Fear	0	0	0	4	1	1	0
Happiness	1	0	0	0	5	0	0
Neutral	0		0	0	0	7	0
Sadness	0	0	0	0	0	0	5

Overall recognition rate = 88.37%

For the first testing, we randomly choose 46 instances form Emo-DB (total 535 instances) with different emotions (anger 8, disgust 5, fear 5, happiness 6, neutral 7, and sadness 5). Then, the rest of the database (491 instances) was used to build a new SVM

base model with the SMO algorithm (named EMO_1 model). The confusion matrix is given in Table 14. In this model, 426 instances were correctly classified, and the overall recognition rate is 86.76%. Then, we use EMO_1 as the base model to test the 46 instances, and the result is shown in Table 15. The average recognition accuracy is 88.37%.

70 instances were randomly chosen from the database IEMOCAP and 10 for each emotion. The rest of the 1027 instances were then used to build the base model (IEMOCAP_1 model) with 988 features using the SMO classification method. The details of this base model are shown in Table 16, in which 650 instances were classified correctly with the average recognition rate at 63.29%. Then, the IEMOCAP_1 is used as the base model for testing the 70 instances. The results were shown in Table 17. Among the 70 instances, 57 correctly result in an 81.42% overall recognition rate.

Table 16

Confusion matrix of the SVM IEMOCAP base model with 1027 instances (988 features)

Labeled	Recognized Emotion						
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	138	5	1	2	4	0	7
Disgust	9	36	11	9	0	4	12
Fear	1	13	38	4	8	9	22
Happiness	5	18	6	73	4	1	18
Neutral	5	1	10	4	133	39	3
Sadness	0	6	11	3	28	143	1
Surprise	14	23	23	17	9	7	89

Overall recognition rate = 63.29%

From the above tests we can see that the average recognition rate of the IEMOCAP database is lower than that of the Emo-DB. The reason is that the speech in the Emo-DB are all acted emotional speech with strong acted emotions, while in the IEMOCAP the speech is most elicited from stories, which means the speech in the IEMOCAP is more close to natural than the Emo-DB.

Table 17

Confusion matrix of the 70 testing instances of the IEMOCAP (988 features)

Labeled	Recognized Emotion						
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	9	0	0	0	1	0	0
Disgust	0	8	1	0	0	1	0
Fear	1	0	8	1	0	0	0
Happiness	1	2	0	7	0	0	0
Neutral	0	0	0	0	10	0	0
Sadness	0	0	1	0	1	8	0
Surprise		0	2	0	1	0	7

Overall recognition rate = 81.42%

Table 18

Confusion matrix of the base model (Emo-DB) with 417 instances for 6 emotions

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	108	1	3	7	1	0
Disgust	1	37	2	0	1	0
Fear	6	1	52	2	1	1
Happiness	10	1	3	51	0	0
Neutral	1	0	0	1	68	1
Sadness	0	1	0	0	0	56

Table 19

Confusion matrix of the IEMOCAP base model with 845 instances for 6 emotions

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	140	11	0	3	3	0
Disgust	10	40	13	13	0	5
Fear	3	18	53	4	7	10
Happiness	6	17	8	88	6	0
Neutral	4	0	6	5	143	37
Sadness	0	5	12	1	27	147

Table 20

Recognition result for 37 instances of the Emo-DB with the Emo-DB base model

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	7	0	0	1	0	0
Disgust	0	4	0	0	1	0
Fear	0	0	4	1	1	0
Happiness	1	0	0	5	0	0
Neutral	0	0	0	0	7	0
Sadness	0	0	0	0	0	7

As we know emotions in the Emo-DB are anger, boredom, disgust, fear, happiness, neutral, and sadness. While emotions in the IEMOCAP are anger, disgust, fear, happiness, neutral, sadness, as well as surprise. In fact, it can recognize more emotions than using a single way by changing some codes in the information fusion. In order to simplify the testing, we just omit the emotion boredom from Emo-DB and the emotion

surprise from IEMOCAP. The detail of the base model for Emo-DB and IEMOCAP are in Table 18 and 19, respectively. The overall recognition rate of the base model for Emo-DB with 6 emotions is 89.21%. While for the IEMOCAP, the accuracy can achieve 72.31%. Both of them are higher than the base model with 7 emotions. The results for the testing instance from the same database are shown in Table 20 and 22 with the recognition rate of 86.49% and 83.33%, respectively. We also use the Emo-DB base model to test the instances from the IEMOCAP and vice versa. The results are shown in Table 21 and 23. From the results we can see that when using the German database Emo-DB to test the English speech most of the instances were recognized as anger, and the recognition rate is only 21.62%. While using the IEMOCAP as the base model most of the instances of Emo-DB were recognized as with a poor recognition rate 23.33%. These results indicate that emotions are language dependent. Thus, it is difficult to generalize from one language to another when doing the speech emotion recognition.

Table 21

Recognition result for 37 instances of the Emo-DB with the IEMOCAP base model

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	8	0	0	0	0	0
Disgust	5	0	0	0	0	0
Fear	5	1	0	0	0	0
Happiness	6	0	0	0	0	0
Neutral	5	2	0	0	0	0
Sadness	5	0	0	0	0	0

Table 22

Recognition result for 70 instances of the IEMOCAP with the IEMOCAP base model

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	9	0	0	0	1	0
Disgust	0	8	1	0	0	1
Fear	1	0	8	1	0	0
Happiness	1	2	0	7	0	0
Neutral	0	0	0	0	10	0
Sadness	0	0	1	0	1	8

Table 23

Recognition result for 70 instances of the IEMOCAP with the Emo-DB base model

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	1	0	0	0	0	9
Disgust	0	3	0	0	0	7
Fear	0	0	0	0	0	10
Happiness	1	2	0	0	2	5
Neutral	0	0	0	0	0	10
Sadness	0	0	0	0	0	10

Testing of Speech Emotion Recognition

Though the database IMOCAP used in the base model is an audio-visual emotional database, the video in this database includes two people, and most of them are side faces. While our multi-sensory emotion recognition system needs frontal faces. Thus,

videos from this database are not suitable for testing the system. Furthermore, to show the objectives of this proposed framework, different data should be tested.

According to our research, the eNTERFACE'05 EMOTION database (Martin et al., 2006) was chosen to test the multi-sensory emotion recognition system. It is an audio-visual emotional database which contains 42 subjects from 14 different countries. All audio files are performed in English. The emotion expressed in this database is the reaction to some specific stories and is elicited emotion. Since emotion is language dependent and the tested database is English, we only used the IEMOCAP to train the base model in the testing.

Table 24

Confusion matrix of the IEMOCAP base model with 892 instances for 6 emotions

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	143	12	0	4	0	8
Disgust	11	41	11	7	5	16
Fear	3	19	41	7	15	20
Happiness	5	15	6	88	2	19
Neutral	0	7	12	1	177	5
Sadness	15	22	26	11	11	107

Overall recognition rate = 66.93%

In the eNTERFACE'05 EMOTION database, the audios are combined with the videos. So, the *FFmpeg* is used to separate the audios from videos. There are 6 emotions in this database – anger, disgust, fear, happiness, sadness, and surprise. We randomly chose 50 instances for each emotion from 44 subjects. The number of the total instances is 300 (including 50 anger, 50 disgust, 50 fear, 50 happiness, 50 sadness, and 50 surprise).

We also adjust the base model to 6 emotions. The base model is an SMO model with 988 features generated from the IEMOCAP database. The confusion matrix is shown in Table 24. The testing results from Weka are in Table 25. 170 instances were recognized correctly from the 300 instances. We also classify the instances with the SVM classifier built-in the OpenSMILE.

Table 25

Emotion recognition results for eNTERFACE'05 EMOTION (OpenSMILE and Weka)

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	39	1	2	7	2	1
Disgust	13	21	2	4	3	7
Fear	4	8	22	4	7	5
Happiness	7	3	3	30	5	2
Sadness	1	6	4	3	33	3
Surprise	6	4	6	9	0	25

Overall recognition rate = 56.67%

Table 26

Emotion Recognition Results for eNTERFACE'05 EMOTION (OpenSMILE)

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	36	4	1	7	0	2
Disgust	12	24	2	5	0	7
Fear	4	12	21	6	4	5
Happiness	9	2	4	27	3	5
Sadness	3	2	5	1	39	0
Surprise	6	5	9	6	2	22

Overall recognition rate = 56.06 %

The sentences are segmented into turns in this process. The results show that there is only a little bit difference between the two approaches, and the recognition rates are 56.67% and 56.06%, respectively (see Table 25 and Table 26).

Facial Emotion Recognition from Videos

In this section, videos from eNTERFACE'05 EMOTION (Martin et al., 2006) database which are matched to the audio files in the speech emotion recognition testing are used. Figure 31 shows some screenshots for the facial emotion recognition test. Table 27 is the summary of the facial emotion recognition results from the videos. The overall recognition rate is 68.67%. The reason for the low recognition rate is that all the instances used here are from a general database, and each instance represents a specific emotion. Thus, there are no specific normal states for initialization.

Figure 31. Screenshot for facial emotion recognition testing.

Table 27

Facial emotion recognition result of the eNTERFACE'05 EMOTION Database

Labeled Emotion	Recognized Emotion						
	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	35	7	1	2	3	2	0
Disgust	7	32	3	2	1	4	1
Fear	2	2	33	2	1	0	10
Happiness	1	1	2	43	2	0	1
Sadness	5	1	5	1	3	33	2
Surprise	1	0	11	2	6	0	30

Overall recognition rate = 68.67%

While our facial emotion recognizer needs neutral state at the beginning of each instance. Thus, there is some bias in the initialization step which results in lower recognition rate.

Table 28

Information fusion without adding weight

Labeled Emotion	Recognized Emotion					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	34	6	0	6	1	3
Disgust	11	36	1	0	2	0
Fear	1	8	30	2	0	9
Happiness	8	2	3	41	2	4
Sadness	2	0	7	0	40	1
Surprise	4	0	13	2	0	31

Overall recognition rate = 70.67%

Results of Information Fusion

Two types of information fusion are tested in this section. Table 28 and 29 show the confusion matrices of the information fusion with and without adding weight. The recognition rates are 70.67% and 74.33%, respectively.

Table 29

Information fusion with adding weight

Labeled	Recognized Emotion					
Emotion	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	39	4	0	4	1	2
Disgust	10	36	1	1	2	0
Fear	0	7	34	2	0	7
Happiness	3	0	1	45	0	1
Sadness	4	1	7	0	37	1
Surprise	5	2	11	1	0	32

Overall recognition rate = 74.33%

Table 30

The detailed accuracy of emotion recognition

	Speech information	Facial expression	Fusion w/o weight	Fusion w weight
Anger	72.00%	70.00%	68.00%	78.00%
Disgust	48.00%	64.00%	72.00%	72.00%
Fear	40.38%	66.00%	60.00%	68.00%
Happiness	54.00%	86.00%	82.00%	90.00%
Sadness	78.00%	66.00%	80.00%	74.00%
Surprise	44.00%	60.00%	62.00%	64.00%
Overall Rec Rate	56.06%	68.67%	70.67%	74.33%

Through information fusion, the improvement of emotion recognition can achieve 18.27% compared to the speech emotion recognition and 5.66% compared to the facial emotion recognition. Figure 32 shows the detail comparison of different emotion recognition. From the figure we can see that adding weight to the information fusion has better results for most of the emotions. In addition, the overall performance is better than fusion without adding weight. Though the speech emotion recognition results are poor, they still contribute to the whole framework through information fusion. Table 30 is the detail information of Figure 32.

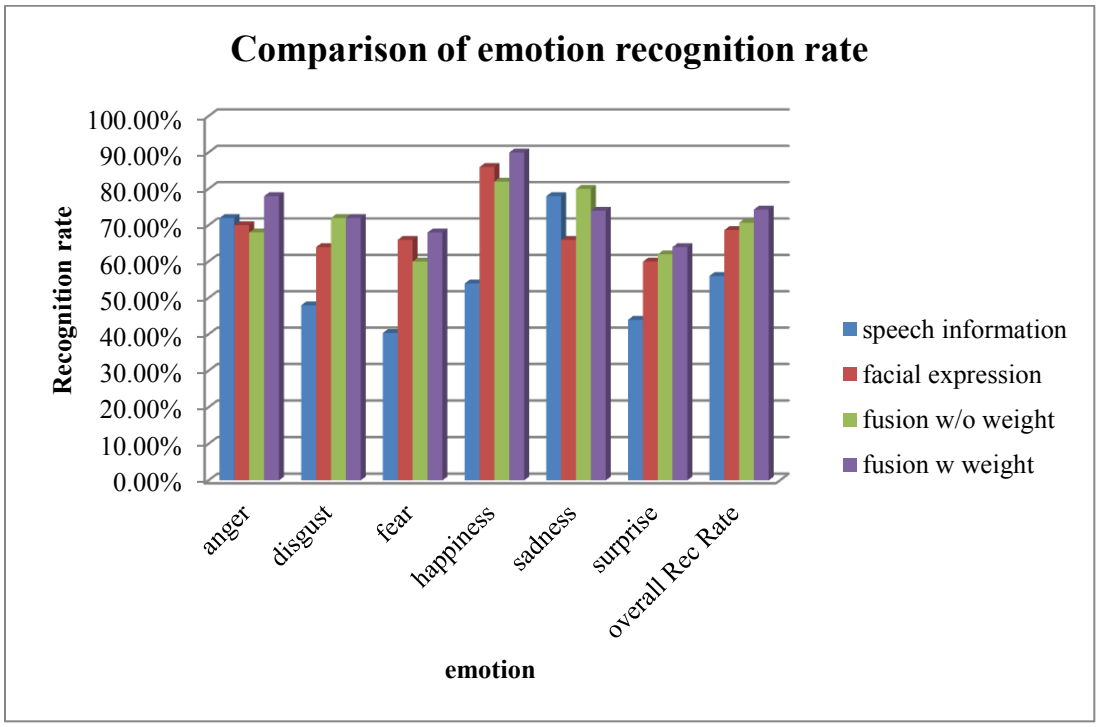

Figure 32. The comparison of multi-sensory emotion recognition.

Facial Emotion Recognition from Images

In this section, the Japanese Female Facial Expression (JAFFE) Database (Lyons, Kamachi, & Gyoba, 1997) was used in testing the facial emotion recognition from images. It contains 213 images from 10 Japanese females with 7 facial expressions (6

basic emotions and 1 neutral state). The JAFFE images can be used for non-commercial research and downloaded from their website freely. Figure 33 is an example of the emotional images of the JAFFE.

In the testing, we first choose 128 images, about 60% of the database, in the training step. Then, the rest 85 images are used to test the recognition accuracy of the proposed method. The images in the training part are randomly chose and at least one for each emotion of each subject. The details of the training data set and testing data set are shown in Table 31. The result is shown in Table 32. The overall recognition rate is 81.59% in this case.

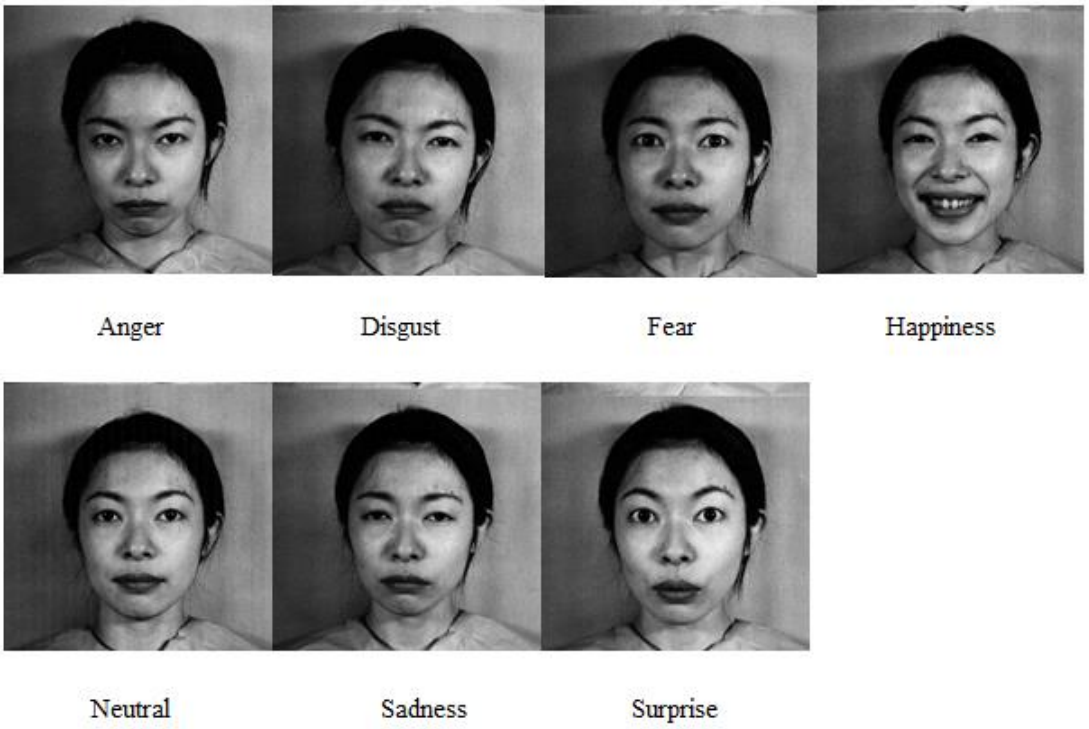

Figure 33. Images of Facial expression in the JAFEE database (Lyons et al., 1997).

Then, we choose 170 images for the training purpose which is about 80% of the JAFEE database. While the rest 20% of the database is utilized for the testing purpose. Table 33 shows the detail of the instance distribution. The testing results are shown in

Table 34 where we can see that the overall recognition rate is 83.54%. This is about 2% improvement compared to the 60%-40% distribution. This result indicates that the more the training data, the better the performance of the testing data.

Table 31

Test Instances detail (60% training and 40% testing)

	Training data	Testing data
Anger	15	15
Disgust	16	13
Fear	18	14
Happiness	21	10
Neutral	19	11
Sadness	19	12
Surprise	20	10

Table 32

Facial emotion recognition from images (60% training and 40% testing)

Labeled	Recognized Emotion						
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	80.00%	0	6.67%	0	0	13.33%	0
Disgust	7.69%	76.92%	7.69%	0	0	7.69%	0
Fear	0	0	85.71%	0	0	14.29%	0
Happiness	0	10.00%	0	90.00%	0	0	0
Neutral	0	0	0	9.09%	81.82%	9.09%	0
Sadness	0	8.33%	16.67%	0	0	66.67%	8.33%
Surprise	0	0	0	0	0	10.00%	90.00%

The overall recognition rate = 81.59%

Table 33

Test Instances detail (80% training and 20% testing)

	Training data	Testing data
Anger	24	6
Disgust	23	6
Fear	25	7
Happiness	24	7
Neutral	25	5
Sadness	25	6
Surprise	24	6

Table 34

Facial emotion recognition from images (80% training and 20% testing)

Labeled	Recognized Emotion						
Emotion	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	100.00%	0	0	0	0	0	0
Disgust	0	83.33%	0	0	0	16.67%	0
Fear	14.29%	0	71.43%	0	0	0	14.29%
Happiness	0	0	0	100.00%	0	0	0
Neutral	0	20.00%	0	0	80.00%	0	0
Sadness	0	16.67%	0	0	0	83.33%	0
Surprise	0	16.67%	16.67%	0	0	0	66.67%

The overall recognition rate = 83.54%

CHAPTER VI

CONCLUSION AND FUTURE WORK

This dissertation has established a multi-sensory emotion recognition system by integrating multi-channel information about speech and facial expression. The goal of the proposed system is to improve emotion recognition accuracy during human computer interaction.

Conclusion

In this dissertation, a multi-sensory emotion recognition framework was proposed which mainly contains three parts: the speech emotion recognizer, the facial emotion recognizer, and information fusion. The proposed approach tries to distinguish between six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) as well as the neutral state by integrating information from audio (speech) and video (facial expression) channels. The proposed framework can also do either speech emotion recognition or facial expression recognition individually.

OpenSMILE is used in speech emotion recognition to process the audio file, extract acoustic features, and classify the emotions. Two emotional databases Emo-DB and IEMOCAP were used to find the base model. We have tested two types of features 988 acoustic features and 384 acoustic features. Experiments demonstrate that models with 988 features have better performance than that of 384 features. Two types of classification algorithm were used. One is the SVM SMO algorithm, and the other one is the Naïve Bayes algorithm. The Naïve Bayer algorithm has better performance only for two types of emotions (fear and neutral has better performance when using 988 features to the IEMOCAP), while the SMO algorithm has a better accuracy in the overall

performance for both Emo-DB and IEMOCAP databases. Thus, in our proposed framework 998 features were chosen in the base model with the SVM SMO classification algorithm. The recognition accuracy of this chosen base model can achieve 88.6% and 64.4% for Emo-DB and IEMOCAP, respectively. The experiment of the speech emotion recognition also indicates that emotion is language dependent. It is hard to have a general model for speech emotion recognition with a different language.

Two approaches were achieved for the facial emotion recognition. One is used to recognize facial emotion from static images. In this approach, skin color segmentation was used for the face detection and Bezier Curve was used for the feature extraction. The JAFFE database was used to test the accuracy of this approach. And the accuracy can achieve to 83.54% when using 80% for the training purposes and 20% for testing. The experiments demonstrate that the more the training data, the better accuracy it can achieve. Another approach is recognizing emotion from recorded video or live video via a webcam. The ASM tracker was used for 116 facial landmarks. Seven high level regions were extracted from the 116 facial landmarks as the feature set. The nearest distance is applied as the classification method. Though the training model is better if it is person-specific, a generic model was built for interacting with different subjects. For the best use of the facial emotion recognizer, one may first train a person-specific model and then use it to recognize emotions.

The instances of the eNTERFACE'05 EMOTION database are used to test the multi-sensory emotion recognition framework. In the speech emotion recognition tests, the recognition accuracy can achieve 56.67% and 56.07% offline with Weka and online with OpenSMILE, respectively. The average facial expression recognition rate is 68.67%.

There are two types of methods used in the information fusion. One fuses the speech and facial emotion recognition results directly. The improvement of the recognition is 2% compared to the facial expression recognition and 14.61% compared to the speech emotion recognition. Another method adds weight to both speech and facial expression results according to Albert Mehrabian's communication model (Mehrabian, 1971), and the recognition accuracy can achieve 74.33%, which is a 3.66 % improvement compare to the infusion without adding weight. It is the first time to add weight in the process of information fusion in this area.

One thing to be pointed out is that, though the result of the facial emotion recognition is much higher than that of the speech, it can achieve even higher result if the test instance is much closer to our requirement. Our requirement for the facial video is that at the beginning, the testing subject should have a neutral state for the initialization. However, most of the test instances do not satisfy this requirement, and we can only use the mean shape of the first m frame (m is the initialization threshold) of the test subject to initialize. Thus, new audio-visual emotional database is needed to demonstrate the advantage of the proposed system in the future.

Future Work

Based on our research results in this dissertation, more opportunities for future research in this field can be done to extend this work. First, more emotional data samples can be collected from a large number of people, and these samples can be used in the speech and facial emotion recognition for the training purpose or testing purpose and also to form an emotional database to convenient other researchers in this area. If such a database can be collected, it may become a milestone in the field of emotion recognition.

Second, it is possible to add feedback to the multi-sensory emotion recognition framework to improve the recognition accuracy by assiging a dynamic weight to the information fusion instead of using the fixed weight (see Figure 34). Another possible extension is to integrate more modalities into the system such as body gesture, hand gesture, etc. Adding more modalities can provide more information about the emotional states of human being, thus, may achieve better emotion recognition accuracy during human computer interaction. Finally, different acoustic features can be tested to improve the robustness of the base model in speech emotion recognition.

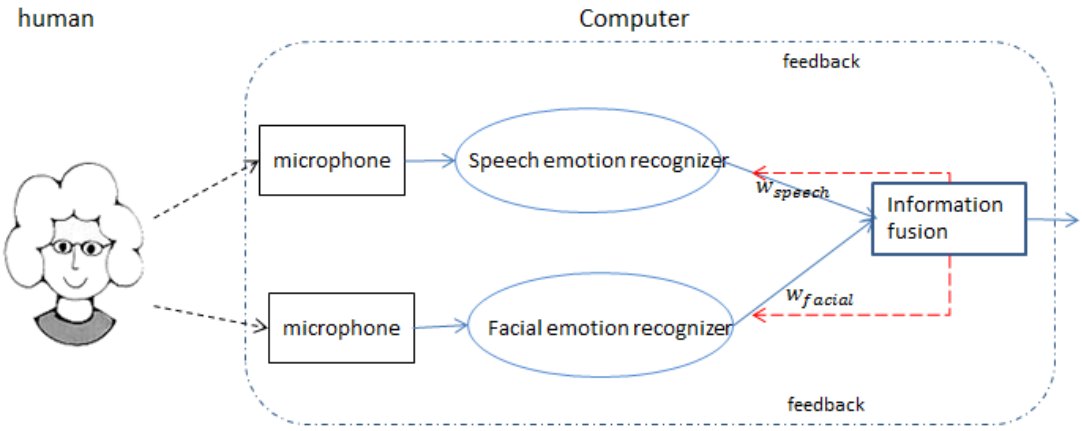

Figure 34. Multi-sensory emotion recognition with feedback.

The multi-sensory emotion recognition framework proposed in this dissertation can be extended to many other research topics in the field of human computer interaction. We hope our research can trigger more investigations in the area of human computer interaction to make computers more friendly, natural, and human-like in the near future.

REFERENCES

Aastha, J. (2013, August). Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering, 3*(8), 387-393.

Baltiner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., & Amir, N. (2011). Whodunnit: searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language. 25*(1), 4-28.

Barra-Chicote, R., Fernandez, F., Lufti, S., Lucas-Cuesta, J., Macias-Guarasa, J., Montero, J., & Pardo, J. (2009). Acoustic emotion recognition using dynamic Bayesian networks and multi-space distributions. *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 336-339). Brighton, UK.

Batliner, A., Hacker, C., Steidl, S., Noth, E., D'Arcy, S., Russell, M., & Wong, M. (2004). You stupid tin box - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. *Proceedings of the 4th International Conference of Language Resources and Evaluation LREC 2004*, (pp. 171-174). Lisbon, Portugal.

Batliner, A., Steidl, S., Seppi, D., & Schuller, B. (2010). Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human-Computer Interaction*, *1*(3), 1-15.

Batliner, A., Zeißler, V., Frank, C., Adelhardt, J., Shi, R., & Nöth, E. (2003). We are not amused — But how do you know? User states in a multi-modal dialogue system.

Proceedings of the 8th European Conference on Speech Communication and Technology (INTERSPEECH), (pp. 733-736). Geneva, Switzerland.

Berlin Database of Emotional Speech. (2014). Retrieved from http://pascal.kgw.tu-berlin.de/emodb/index-1280.html.

Black, M., & Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. *In Proceedings of the Fifth International Conference on Computer Vision*, (pp. 347-381). Cambridge, MA.

Bozkurt, E., Erzin, E., Erdem, C., & Erdem, A. (2010). Interspeech 2009 emotion recognition challenge evaluation. *Signal Processing and Communications Applications Conference (SIU), 2010 IEEE 18th*, (pp. 216-219). Diyarbakir, Turkey.

Bruce, V. (1992). What the Human Face Tells the Human Mind: Some Challenges for the Robot-Human Interface. *Proceedings. IEEE International Workshop on Robot and Human Communication*, (pp. 44-51). Tokyo, Japan.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. *In Proceedings of Interspeech*, (pp. 1517-1520). Lissabon, Portugal.

Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation, 42*(4), 335-359.

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., & Karpouzis, K. (2007). Multimodal emotion recognition from expressive faces,

body gestures and speech. *Artificial Intelligence and Innovations 2007: From Theory to Applications, 247*, 375-388.

Casale, S., Russo, S., Scebba, G., & Serrano, S. (2008). Speech emotion classification using machine learning algorithms. *Proceedings of the IEEE International Conference on Semantic Computing*, (pp. 158-165). Santa Clara, CA.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent System and Technology, 2*(3), 1-27.

Cho, J., Kato, S., & Itoh, H. (2007). Bayesian-based inference of dialogist's emotion for sensitivity robots. *Proceedings of the 16th IEEE International Conference on Robot & Human Interactive Communication*, (pp. 792–797). Jeju Island, South Korea.

Cichosz, J., & Slot, K. (2007). Emotion recognition in speech signal using emotion-extracting binary decision trees. *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII): Doctoral Consortium*, (pp. 9-16). Lisbon, Portugal.

Clavel, C., Devillers, L., Richard, G., Vasilescu, I., & Ehrette, T. (2007). Detection and analysis of abnormal situations through fear-type acoustic manifestations. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (pp. IV - 21-24). Honolulu, Hawaii.

Cohen, I., Garg, A., & Huang, T. (2000). Emotion Recognition from Facial Expressions using Multilevel HMM. *In Neural Information Processing Systems Workshop on Affective Computing,* Denver, CO

Colmenarez, A., & Huang, T. (1997). Face Detection with Information-Based Maximum Discrimination. *Proceedings, IEEEcomputer Society Conference on Computer Vision and Pattern Recognition*, (pp. 782-787). San Juan, Pueto Rico.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(6), 681-685.

Cootes, T. F., Taylor, C. J., Cooper, D., & Graham, J. (1995). Active Shape Models - Their Training and Application. *Computer Vision and Image Understanding, 61*(1), 38-59.

Corradini, A., Mehta, M., Bernsen, N. O., Martin, J. C., & Abrilian, S. (2003). Multimodal input fusion in human-computer interaction on the example of nice project. *Proceedings of the NATO*, 223-234.

Dam, A. (2000). Beyond WIMP. *IEEE Computer Graphics and Applications*, (pp. 50-51). Los Alamitos, CA.

Darwin, C. (1998). *The expression of the emotions in man and animals.* New York, NY, Oxford University Press.

De Silva, L., & Hui, S. (2003). Real-time facial feature extraction and emotion recognition and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on *Information, Communications and Signal Processing*, (pp. 1310-1314). Singapore.

De Silva, L., & Pei, C. (2000). Bimodal emotion recognition. *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, (pp. 332-335). Grenoble, France.

Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion in speech. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, (pp. 1970-1973). Philadelphia, PA.

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks, 18*(4), 407-422.

Donato, G., Bartlett, M., Hager, J., Ekman, P., & Sejnowski, T. (1999). Classifying Facial Actions. *IEEE Trans. Pattern Analysis and Machine Intelligence, 21*(10), 974-898.

Ekman, P. (1982). *Emotion in the human face.* Cambridge, UK: Cambridge University Press.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*, *6*(3-4), 169-200.

Ekman, P. (1999). Basic emotions. In *Handbook of cognition and emotion* (pp. 45-60). Sussex, UK: John Wiley & Sons, Ltd.

Ekman, P., & Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*, 124-129.

Ekman, P., & Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto, CA: *Consulting Psychologists Press.*

Engberg, I., & Hansen, A. (1996). *Documentation of the Danish Emotional Speech Database (DES).* Technic Report, Aalborg University, Aalborg, Denmark.

Essa, I., & Pentland, A. (1997). Coding, Analysis, Interpretation, and Recognition of Facial Expressions. *IEEE Transactions, Pattern Analysis and Machine Intelligence, 19*(7), 757-763.

Eyben, F., Wollmer, M., & Schuller, B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the international conference on ACM Multimedia (MM'10), ACM*, (pp. 1459-1462). Firenze, Italy.

Fasel, B., & Luettin, J. (2003, January). Automatic facial expression analysis: a survey. *Pattern Recognition, 36*(1), 259-275.

Fernandez, R., & Picard, R. (2003). Modeling drivers' speech under stress. *Speech Communication, 40*(1-2), 145-159.

Fernandez, R., & Picard, R. (2005). Classical and novel discriminant features for affect recognition from speech. *Proceedings of the 9th European Conference on Speech Communication and Technology*, (pp. 473-476). Lisbon, Portugal.

FFmpeg. (2012, 01 04). Retrieved from FFmpeg: http://www.ffmpeg.org/about.html

Friesen, W., Ekman, P., & Hager, J. (2002). The facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologist.* San Francisco, CA.

Graf, H., Chen, T., Petajan, E., & Cosatto, E. (1995). Locating Faces and Facial Parts. *Proceedings of the First International Workshop on Automatic Face and Gesture Recognition*, (pp. 41-46). Zurich, Switzerland.

Grandjean, D., Sander, D., & Scherer, K. R. (2008, June 01). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness & Cognition, 17*(2), 484-495.

Grimm, M., Mower, E., Kroschel, K., & Narayanan, S. (2006). Combining categorical and primitives-based emotion recognition. *Proceedings of the 14th European Signal Processing Conference.* Florence, Italy.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1).

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences, 16*(9), 458-466.

Hargrove, L. J., Li, G., Kevin, E. B., & Bernie, H. S. (2009). Principal Components Analysis Preprocessing for Improved Classification Accuracies in Pattern-Recognition-Based Myoelectric Control. *Biomedical Engineering, IEEE Transactions, 56*(5), 1407-1414.

Hassan, A., & Damper, R. (2009). Emotion recognition from speech using extended feature selection and a simple classifier. *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 2043-2046). Brighton, UK.

IEMOCAP DATABASE. (2014). Retrieved from http://sail.usc.edu/iemocap/

Johnstone, T., & Scherer, K. R. (2000). Vocal Communication of Emotion. In M. Lewis, & J. Haviland, *The Handbook of Emotion* (pp. 220-235). New York, NY: Guilford.

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Fourth IEEE International Conference: Automatic Face and Gesture Recognition* (pp. 46-53). Grenoble, Russia: IEEE.

Kim, S.-H., & Ahn, Y. (2007). An Approximation of Circular Arcs by Quartic Bezier Curves. *Computer-aided Design, 39*(6), 490-493.

Kwon, O.-W., Chan, K., Hao, J., & Lee, T.W. (2003). Emotion recognition by speech signals. *Proceedings of the 8th European Conference on Speech Communication and Technology (INTERSPEECH)*, (pp. 125-128). Geneva, Switzerland.

Kwon, Y. H., & da Vitoria Lobo, N. (1994). Face Detection Using Templates. *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, (pp. 764-767). Jerusalem, Israel.

Lanitis, A., Taylor, C., & Cootes, T. (1995). A unified approach to coding and interpreting face images. *Proceedings of the 5th International Conference on Computer Vision (ICCV)*, (pp. 368-373). Cambridge, UK.

Lanitis, A., Taylor, C., & Cootes, T. (1995). An Automatic Face Identification System Using Flexible Appearance Models. *Image and Vision Computing, 13*(5), 393-401.

Lee, C., & Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing, 13*(2), 293-303.

Lee, C., Yildirim, S., Bulut, M., & Kazemzadeh, A. (2004). Emotion recognition based on phoneme classes. *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH)*, (pp. 889–892). Jeju Island, South Korea.

Leon, J., & Pantic, R. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1424-1445.

Leung, T., Burl, M., & Perona, P. (1995). Finding Faces in Cluttered Scenes Using Random Labeled Graph Matching. *Proceedings of the Fifth IEEE International Conference on Computer Vision*, (pp. 637-644). Cambridge, MA.

Leung, T., Burl, M., & Perona, P. (1998). Probabilistic Affine Invariants for Recognition. *Proceeding of 1998. IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 678-684). Santa Barbara, CA.

Lew, M. (1996). Information Theoretic View-Based and Modular Face Detection. *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, (pp. 198-203). Killington, VT.

Lugger, M., & Yang, B. (2008). Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (pp. 4945-4948). Las Vegas, NV.

Lyons, M. J., Kamachi, M., & Gyoba, J. (1997). Japanese Female Facial Expressions (JAFFE). *Database of digital images*, Retrieved from http://www.kasrl.org/jaffe_info.html.

Lyons, M., Budynek, J., Plante, A., & Akamatsu, S. (2000). Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, (pp. 202-207). Grenoble, French.

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 Audio-Visual Emotion Database. *Proceedings of the 22nd International Conference on Data Engineering Workshops.* Mons, Belgium: IEEE Computer Society.

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions*, *74*, 3474-3483.

Mehrabian, A. (1971). *Silent Message (1st ed.).* Belmont: Wadsworth.

Meyers, D. G. (2004). Theories of Emotion. In *Psychology: Seventh Edition.* New York, NY: Worth Publishers.

Mower, E., Mataric, M., & Narayanan, S. (2009). Evaluating evaluators: A case study in understanding the benefits and pitfalls of multi-evaluator modeling. *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 1583–1586). Brighton, UK.

Neiberg, D., & Elenius, K. (2008). Automatic recognition of anger in spontaneous speech. *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 2755-2758). Brisbane, Australia.

OpenCV. (2014). Retrieved from OpenCV: http://opencv.org/

Ortony, A., & Turner, T. J. (1990, July). What's basic about basic emotions? *Psychological Review, 97*(3), 315-331.

Ostermann, J. (1998). Animation of synthetic faces in Mpeg-4. *Computer Animation*, (pp. 49-51). Philadelphia, PA.

Osuna, E., Freund, R., & Girosi, F. (1997). Training Support Vector Machines: An Application to Face Detection. *Proceedings of 1997 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 130-136). San Juan, Puerto Rico.

Otsuka, T., & Ohya, J. (1997). Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences.

Proceedings of the International Conference on Image Processing (ICIP-97, (pp. 546-549). Santa Barbara, CA.

Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: Features and algorithms. *International Journal of Human-Computer Studies, 59*(1-2), 157-183.

Pantic, M., & Rothkrantz, L. J. (2000, December). Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1424-1445.

Pantic, M., & Rothkrantz, L. J. (2003). Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE*, *91*(9), 1370-1390.

Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. *Proceedings of 1999 Artificial Neural Networks in Engineering*, (pp. 7-10). New York, NY.

Petrushin, V. (2002). Creating emotion recognition agents for speech signal. In K. Dautenhahn, A. Bond, L. Canamero, & B. Edmonds, *Creating Relationships with Computers and Robots* (pp. 77-84). Norwell: Kluwer Academic Publishers.

Picard, R. (1997). *Affective Computing.* Cambridge, MA: MIT press.

Planet, S., Iriondo, I., Socoró, J.C., Monzo, C., & Adell, J. (2009). GTM-URL contribution to the INTERSPEECH 2009 Emotion Challenge. *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 316-319). Brighton, UK.

Plutchik, R. (1980). Emotion: Theory, research, and experience: Vol.1. In *Theories of emotion 1*. New York, NY: Academic.

PortAudio. (2014). Retrieved from PortAudio: http://www.portaudio.com/download.html

Rajagopalan, A., Kumar, K., Karlekar, J., Manivasakan, R., Patil, M., Desai, U., & Chaudhuri, S. (1998). Finding Faces in Photographs.In *Proceedings of the Sixth IEEE International Conference on Computer Vision*, (pp. 640-645). Washington, DC: IEEE Computer Society.

Rosenblum, M., Yacoob, Y., & Davis, L. (1996). Human expression from motion using a radial basis function network architecture. *IEEE Transactions on Neural Network, 18*, 636-642.

Rowley, H., Baluja, S., & Kanade, T. (1998). Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(1), 23-38.

Sakai, T., Nagao, M., & Fujibayashi, S. (1969). Line Extraction and Pattern Detection in a Photograph. *Pattern Recognition*, *1*(3), 233-248.

Samal, A., & Iyengar, P. (1995). Human Face Detection Using Silhouettes. *International Journal of Pattern Recognition and Artificial Intelligence, 9*(6), 845-867.

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal Processes in Emotion: Theory, Methods, Research.* New York, NY: Oxford University Press.

Schiel, F., Steininger, S., & Turk, U. (2002). The SmartKom multimodal corpus at BAS. *Proceedings of the 3rd Language Resources & Evaluation Conference (LREC)*, (pp. 200-206). Gran Canaria, Canary Islands.

Schneiderman, H., & Kanade, T. (1998). Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (pp. 45-51). Washington, DC.

Schuller, B., Baltiner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., & Aharonson, V. (2007). The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. *Proceedings of Interspeech*, (pp. 2253-2256). Antwerp, Belgium.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognizing realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, *53*(9-10), 1062-1087.

Schuller, B., Lang, M., & Rigoll, G. (2002). Multimodal emotion recognition in audio visual communication. *In Proceedings of 2002 IEEE International Conference on Multimedia and Expo, 2002. ICME'02*, (pp. 745-748). Lausanne, Switzerland.

Schuller, B., Muller, R., Lang, M., & Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Proceedings of Interspeech 2005*, (pp. 805-808). Lisbon, Portugal.

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge. *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 312-315). Brighton, UK.

Sethu, V., Ambikairajah, E., & Epps, J. (2009). Pitch contour parameterisation based on linear stylisation for emotion recognition. *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 2011-2014). Brighton, UK.

Sidorova, J. (2009). Speech emotion recognition with TGI+.2 classifier. *Proceedings of the 12th Conference of the European Chapter of the Association for*

Computational Linguistics (EACL): Student Research Workshop, (pp. 54-60). Athens, Greece.

Sirohey, S. (1993). *Human Face Segmentation and Identification* (CS-TR-3176). University of Maryland.

Sobel operator. (n.d.). Retrieved from http://docs.opencv.org/doc/tutorials/imgproc/imgtrans/sobel_derivatives/sobel_derivatives.html

Soleymani, M., Pantic, M., & Pun, M. (2012). Multimodal emotion recognition in response to videos. *Affective Computing, IEEETransactions on, 3*(2), 211-223.

Sonka, M., Hlavac, v., & Boyle, R. (2008). *Image Processing, Analysis, and Machine Vision, Third Edition.* Tomson Engineering.

Sumpeno, S., Hariadi, M., & Purnomo, M. H. (2011). Facial Emotional Expressions of Life-like Character Based on Text Classifier and Fuzzy Logic. *IAENG International Journal of Computer Science*.

Sung, K.-k., & Poggio, T. (1998). Example-Based Learning for View-Based Human Face Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence, 20*(1), 39-51.

Tomkins, S. S. (1962). *Affect imagery consciousness: the positive affects.* New York, NY: Springer.

Tomkins, S. S. (1963). *Affect imagery consciousness: the negative affects.* New York, NY: Springer.

Tsukamoto, A., Lee, C.-W., & Tsuji, S. (1994). Detection and Pose Estimation of Human Face with Synthesized Image Models. *Proceedings of the 12the International Conference on Pattern Recognition*, (pp. 745-757). Jerusalem, Israel.

Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience, 3*(1), 71-86.

Vereveridis, D., & Kotropoulo, C. (2003). A Review of Emotional Speech Databases. *PCI 2003, 9th PanHellenic Conference on Informatics*, (pp. 560-574). Thessaloniki, Greece.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1161-1181.

Ververidis, D., & Kotropoulos, C. (2006). Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. *Proceedings of the 14th European Signal Processing Conference (EUSIPCO)*. Florence, Italy.

Viola, P., & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple. *Computer Vision and Pattern Recognition, 57*(2), 137-154.

Vlasenko, B., Schuller, B., Mengistu, K., Rigoll, G., & Wendemuth, A. (2008). Balancing spoken content adaptation and unit length in the recognition of emotion and interest. *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 805-808). Brisbane, Australia.

Vogt, T., & Andre, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. *Multimedia and Expo, 2005. ICME 2005 IEEE International Conference*, (pp. 474-477). Amsterdam, Kingdom of the Netherlands.

Vogt, T., & Andre, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. *Proceedings of Language Resources and Evaluation Conference(LREC 2006).* Genoa, Italy.

Vogt, T., AndrÃl', E., & Wagner, J. (2008). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realization. *Lecture Notes in Computer Science on Affect and Emotion in Human-Computer Interaction, 4868*, 75-91.

Vogt, T. (2010). *Real-time automatic emotion recognition from speech*, (Doctoral dissertation). Retrieved from http://pub.uni-bielefeld.de/luur/download?func=downloadFile&recordOId=2301483&fileOId=2301486

Wagner, J., Vogt, T., & Andre, E. (2007). A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech. *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII)*, (pp. 114-125). Lisbon, Portugal.

Walker, W., Lamere, P., Kwok, P., Raj, B., Sigh, R., Gouvea, E., & Woelfel, J. (2004). *Sphinx-4: a flexible open source framework for speech recognition.* Mountain View, CA: Technical report.

Wang, Y., & Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia*, *10*(4), 659-668.

Wei, Y. (2009). *Research on Facial Expression Recognition and Synthesis*. (Master's Thesis, Department of Computer Science and Technology, Nanjing University). Retrieved from http://code.google.com/p/asmlibrary

Wierzbicka, A. (1992). Talking about emotions: Semantics, culture, and cognition. *Cognition and Emotion*, *6*, 285-319.

Witten, I., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques.* San Francisco, CA: Morgan Kaufmann.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Dougla-Cowie, E., & Cowie, R. (2008). Abandoning emotion classes — Towards continuous emotion recognition with modelling of long-range dependencies. *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 597-600). Brisbane, Australia.

Wu, L., Oviatt, S. L., & Cohen, P. R. (1999). Multimodal integration: A statistical view. *IEEE Transactions on Multimedia*, *1*(4), 334-341.

Wu, S., Falk, T., & Chan, W.-Y. (2008). Long-term spectro-temporal information for improved automatic speech emotion classification. *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (pp. 638-641). Brisbane, Australia.

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2007). *Hierarchical classification of emotional speech.* Technical Report, Laboratoire d'Informatique en Image et Systèmes d'Information (LIRIS), Université de Lyon, Lyon, France.

Yacoob, Y., & Davis, L. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*, 636-642.

Yang, G., & Huang, T. (1994). Human Face Detection in Complex Background. *Pattern Recognition, 27*(1), 53-63.

Yang, M.H., Kriegman, D., & Ahuja, N. (2002). Detecting faces in images: a survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions, 24*(1), 34-58.

You, M., Chen, C., Bu, J., Liu, J., & Tao, J. (2007). Manifolds based emotion recognition in speech. *Computational Linguistics and Chinese Language Processing*, *12*, 49-64.

Yow, K., & Cipolla, R. (1886). A probabilistic Framework for Perceptual Grouping of Features for Human Face Detection. *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, (pp. 16-21). Killington, VT.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39-58.

Zhang, C., & Zhang, Z. (2010). *A Survey of Recent Advances in Face Detection.* Redmond, WA: Microsoft Research.

Zhou, J., Wang, G., Yang, Y., & Chen, P. (2006). Speech emotion recognition based on rough set and SVM. *Proceedings of the 5th IEEE International Conference on Cognitive Informatics (ICCI)*, (pp. 53-61). Beijing, China.