Human Emotion Recognition from Body Language of the Head using Soft Computing Techniques

Yisu Zhao

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements

For the degree of Doctorate of Philosophy in Computer Science

Ottawa-Carleton Institute for Computer Science (OCICS)

Faculty of Engineering

University of Ottawa

Abstract

When people interact with each other, they not only listen to what the other says, they react to facial expressions, gaze direction, and head movement. Human-computer interaction would be enhanced in a friendly and non-intrusive way if computers could understand and respond to users' body language in the same way.

This thesis aims to investigate new methods for human computer interaction by combining information from the body language of the head to recognize the emotional and cognitive states. We concentrated on the integration of facial expression, eye gaze and head movement using soft computing techniques. The whole procedure is done in two-stage. The first stage focuses on the extraction of explicit information from the modalities of facial expression, head movement, and eye gaze. In the second stage, all these information are fused by soft computing techniques to infer the implicit emotional states.

In this thesis, the frequency of head movement (high frequency movement or low frequency movement) is taken into consideration as well as head nods and head shakes. A very high frequency head movement may show much more arousal and active property than the low frequency head movement which differs on the emotion dimensional space. The head movement frequency is acquired by analyzing the tracking results of the coordinates from the detected nostril points.

Eye gaze also plays an important role in emotion detection. An eye gaze detector was proposed to analyze whether the subject's gaze direction was direct or averted. We proposed a geometrical relationship of human organs between nostrils and two pupils to achieve this task. Four parameters are defined according to the changes in angles and the changes in the proportion of length of the four feature points to distinguish avert gaze from direct gaze. The sum of these parameters is considered as an evaluation parameter that can be analyzed to quantify gaze level.

The multimodal fusion is done by hybridizing the decision level fusion and the soft computing techniques for classification. This could avoid the disadvantages of the decision level fusion technique, while retaining its advantages of adaptation and flexibility. We introduced fuzzification strategies which can successfully quantify the extracted parameters of each modality into a fuzzified value between 0 and 1. These fuzzified values are the inputs for the fuzzy inference systems which map the fuzzy values into emotional states.

Acknowledgements

First, I would like to express my deep and sincere appreciation to my supervisors Dr. Nicolas D. Georganas and Dr. Emil M. Petriu for their excellent guidance and warm encouragement throughout my Ph.D study. I have benefited greatly from their vision, technical insights and profound thinking. I feel extremely lucky to have studied under their supervision.

I would also like to gratefully acknowledge Dr. Thomas Whalen and Miriam Goubran, who provided me with tremendous help in building the psychological model. Without their expertise in psychology, this dissertation would not have existed.

I would like to thank my journal paper reviewers for their constructive comments and valuable advices. I also wish to thank all my colleagues in Discover Lab for their suggestions and help toward my research.

Last but not least, I want to say thank you to my family for all their love and encouragement. Without their endless support, I would never have made it.

CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Emotion is one of the most complex, psychophysiological, and inherent human behaviours [1]. In daily social life, sensing the emotional state of a human counterpart is crucial. However, when it comes to a computer, is it necessary for it to recognize the emotional state of its counterpart?

With the growing use of computers in modern society, computers are increasingly playing a more important role in many facets of human lives. Traditional human computer interfaces, such as the keyboard and the mouse, ignore implicit information about human users. It is claimed that the trend for computer interfaces is switching from computer-centered designs to human-centered designs, i.e., built to provide a naturally communicative environment between human and computer [2], [5].

Human computer interaction (HCI) can be significantly enhanced if computers acquire the ability to understand and respond to users' needs in a natural, friendly, and non-intrusive manner. To achieve effective human computer intelligent interaction (HCII), a computer must be able to perform human-like interactions with its human

counterpart similar to human-human interactions. Specifically, HCII interfaces must have the ability to detect subtle changes in the user's behaviour, especially the user's emotional states, and to initiate interactions based on this information as opposed to simply responding to the user's commands. For example, a learning environment that is sensitive to emotions and able to detect and respond to students' frustration is expected to increase their motivation and improve their learning abilities as opposed to a system that ignores student emotions [3]. Figure 1.1 shows the evolution of the human computer interface.

Human-human interaction has provided the groundwork to develop automatic ways for computers to recognize emotional expressions as a goal toward achieving HCII. Humans interact with each other to naturally express emotions and feelings by using multimodal means simultaneously so that one complements and enhances another. Researchers have tried to analyze these modalities in an attempt to interpret and categorize emotions for HCII applications. This emerging field has also attracted increasing attention from researchers of diverse fields such as computer science, cognitive science, neuroscience, and related disciplines (see Figure 1.2). Potential applications of emotion detection for HCII cover many areas including gaming, e-learning, security, health care, and elder care.

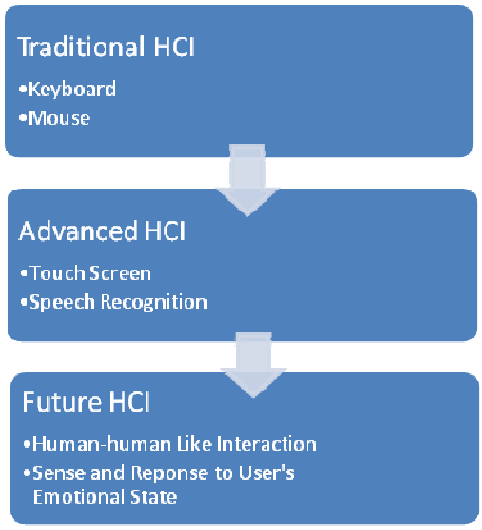


Figure 1.1 The evolution of HCI interface

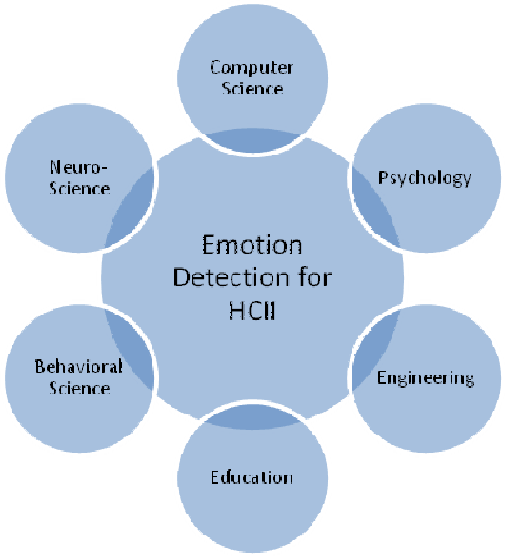


Figure 1.2 Multidisciplinary knowledge of emotion detection for HCII

Human emotion can be expressed through verbal as well as nonverbal cues. Nonverbal information is vital and should be considered and investigated as an indispensable aspect of emotion. Ekman and Friesen [4] give a rationale for studying nonverbal behaviour. They suggest five reasons for studying nonverbal behaviour:

1. Human emotional cues such as intimacy and anxiety are often expressed nonverbally long before they are expressed verbally.
2. Nonverbal expressions often either support or contradict an overt statement of feeling.
3. Bodily expressions sometimes reveal unconscious attitudes about one's affect.
4. Nonverbal behaviours may provide feedback on whether someone is listening, is bored, is getting ready to talk, and so forth.
5. Nonverbal behaviour is probably less affected than verbal behaviour by the censoring of communication.

1.2 Motivation

The publication of *The Expression of the Emotions in Man and Animals* by Charles Darwin in 1872 laid the foundation for later research on nonverbal communication and behaviour. Nonverbal cues are essential in conveying interactive information among humans, mammals, and some other animal species. Figure 1.3 shows an example of the

body language of a cat and a dog [7]. The cat arched its back and raised its hair to show apparent increase in its body size when in the emotion of fear. It makes the cat look larger and more ferocious and decreases the chances of a direct confrontation. Cats and dogs both pulled back their ears and bared their teeth when in the mixed emotion of fear and anger. Such adaptive behaviour has been observed in many species. Emotional body language serves important functions in the lives of animals. They act as signals and as preparations for action. They communicate information from one animal to another about what is likely to happen and thereby affect the chances of survival. Anything that increases the chances of survival tends to be maintained during the course of evolution [7]. Since Darwin's publication in 1872, a lot of research has been conducted regarding the topic of nonverbal communication and behaviour from body language.

Figure 1.3 Examples of animal body languages [7]

Human body language is a form of nonverbal cues that is used in human-human interactions almost entirely subconsciously. Researchers stated a dominate 93% of human communication is achieved through body language, while a merely 7% of communication is consisted of actual words [24], [126]. Moreover, body language can convey emotional state of a person in human communication such as depression, enjoyment, boredom, etc. Body language can also enhance verbal communication. For example, eye contact and nodding give clue to the person you are interacting with that you understand the subject.

The human head is a multi-signal communicative media with remarkable flexibility and specificity [105]. When communicating, crucial information such as personal identification, ethnicity, gender, and age, attractiveness can be gathered from the human

head. Psychological researchers have found that body language of the head can provide significant information for detecting and interpreting emotions [127], [128], [129]. It can convey both explicit and implicit information of one's emotional state and intentions through multi-channel modalities. These important channels include facial expression, eye gaze, head movement, and so on.

This thesis focused on detecting emotional states from the body language of the head using computer vision techniques and soft computing techniques. OpenCV and Matlab are used to implement the proposed methods and evaluate the results.

The remainder of this chapter is organized as follows: in Section 1.2, we discuss the problem statement and the major challenges. Section 1.3 and Section 1.4 outline the contributions made in this thesis and the related publications, respectively. Finally, the organization of this thesis is laid out in Section 1.5.

1.3 Problem Statement

Despite the wide range of modalities, most researchers have focused on facial expression recognition to detect emotion. This has been based on the assumption that expression is triggered for a short duration when the emotion is experienced [20]. Recently, the advantage of multimodal human computer interaction systems has been recognized and many researchers are attempting to integrate different sensing modalities to improve the performance of affective behaviour recognition.

The aim of this thesis is to explore new solutions for human emotion analysis and interpretation. The new exploration and major challenges for emotion detection include the following:

- Is the inferred emotion always congruent with the facial expression? When a facial expression is accompanied by other nonverbal cues, entirely different emotions may be inferred. For example, the smiling facial expression usually means happiness. However, if the person is smiling while shaking his or her head at the same time, this may mean that he or she is not happy. In this case, the head movement is the more important variable in emotion recognition and negates the facial expression. Therefore, considering other modalities is crucial for analyzing emotions.
- Does the head nodding and shaking frequency correlate with human emotion? Head movement not only can provide information of the movement direction (head nod

or head shake), but also can differ in speed and frequency (high frequency movement or low frequency movement). For example, a very high frequency head nodding may show much more arousal and active property than the low frequency head nodding which differs on the emotion dimensional space.

- Does eye gaze only provide information such as attentive or non-attentive and could it provide additional information to aid with emotion recognition? In past research, eye gaze was considered in human behavioural recognition (e.g., the detection of human fatigue and attentive states). However, gaze direction also has a strong relationship with emotion. Psychologists find that a direct gaze is more likely to be associated with approach-oriented emotion, while an averted gaze is more likely to be associated with avoidance-oriented emotion. These findings suggest that gaze direction influences the processing of emotion displays.
- How should the modalities be fused to infer emotional states? Is there any other solution other than the traditional fusion techniques? There is no crisp boundary between high frequency movement and low frequency movement or direct gaze and averted gaze. Therefore, how to classify these modalities is also a major concern of the thesis.
- Is there a cultural influence when expressing emotions? Obtaining the ground truth has always been a big issue in all types of human behavioural detection. However, emotions also have "accent." People from different cultures can display the same emotion in totally different ways. Therefore, cultural differences need to be considered when acquiring the ground truth.

Although the questions noted above have not been taken into consideration in the design of current emotion detection systems, they are important factors and can enrich current emotion detection systems. Based on the above questions, all these factors are taken into consideration in this thesis and solutions are provided for each of them.

1.4 Contributions

The goal of this thesis is to explore new solutions for HCII by showing how to integrate information from the body language of the head to infer emotional and cognitive

states by concentrating on the combination of facial expression, eye gaze and head movement. The following summarizes the contributions of the thesis:

- ✧ A two-stage approach is proposed. The first stage analyzes the explicit information from the modalities of facial expression, head movement, and eye gaze separately. In the second stage, all the information is fused to infer implicit secondary emotional states. By integrating the channels from the body language of the head, the distinguished emotion may result in a different quadrant in the emotional dimension space compared to the corresponding facial expression.

- ✧ To analyze head movements, not only the direction but also the frequency is observed. The head movement frequency is measured by analyzing the tracking results of the coordinates from the detected nostril points. Emotional states under five states of head movement include high frequency head nodding, low frequency head nodding, still, low frequency head shaking, and high frequency head shaking are analyzied.

- ✧ Eye gaze direction is integrated with other head information to analyze emotional states. A geometrical relationship of human organs between the nostrils and the two pupils is developed to achieve this task. Four parameters are defined according to the changes in angles and the changes in the proportion of the length of the four feature points to distinguish averted gaze from direct gaze. The sum of these parameters is considered an evaluation parameter that can be analyzed to quantify gaze level.

- ✧ New solutions are explored for multimodality fusion by hybridizing the decision level fusion and the soft computing techniques to infer emotions. This could avoid the disadvantages of the decision level fusion technique while retaining its advantages of adaptation and flexibility.

- ✧ A fuzzification strategy is proposed which can successfully quantify the extracted parameters of each modality into a fuzzified value between 0 and 1. These fuzzified values are the inputs for the fuzzy inference systems that map the fuzzy values into emotional states.

1.5 Publications Arising from the Thesis

The following refereed book chapter, journal and conference publications have arisen from the work presented in this thesis.

Refereed Book Chapter

1. Y. Zhao, M.D. Cordea, E. M. Petriu, and T. E. Whalen, "Multimedia-Based Affective Human-Computer Interaction," in *Multimedia Image and Video Processing,* 2nd ed., L. Guan, Y. He, S. Y. Kuang, Eds. Boca Raton: CRC Press, 2012, pp. 173-194.

Refereed Journal

1. Y. Zhao, X. Wang, M. Goubran, T. E. Whalen, and E. M. Petriu, "Human Emotion and Cognition Recognition from Body Language of the Head Using Soft Computing Techniques," (accepted) *Journal of Ambient Intelligence and Humanized Computing,* Springer Berlin/Heidelberg, Feb. 2012.

Refereed Conference Proceedings

1. Y. Zhao, X. Wang, and E. M. Petriu, "Facial Expression Analysis Using Eye Gaze Information," *in Proc. IEEE CIMSA'11- Computational Intelligent for Measurement Systems and Applications,* Sept. 2011.

2. Y. Zhao, X. Shen, and N. D. Georganas, "Facial Expression Recognition by Applying Multi-step Integral Projection and SVMs," in *Proc. IEEE I2MTC'09 - Instrumentation and Measurement Technology Conference,* 2009.

3. Y. Zhao, X. Shen, and N. D. Georganas, "Combining Integral Projection and Gabor Transformation for Automatic Facial Feature Detection and Extraction," in *Proc. IEEE HAVE'08 – Haptic Audio Visual Environments and Games,* 2008.

1.6 Thesis Organization

The remainder of this dissertation proceeds as follows:

Chapter 2 seeks to introduce and discuss issues on the approaches of emotion categories, single emotion detection channels, multi-channel emotion detection, and the fusion techniques.

Chapter 3 examines all of the extracted information and considers as input data for soft computing techniques to infer emotional and cognitive states. The fuzzy rules were defined based on the opinion of an expert in psychology, a pilot group, and annotators. In this work, we utilize both fuzzy logic and neural fuzzy techniques for our emotional model test.

Chapter 4 develops a simple, fast, and effective automatic head movement detection system. Nostrils are detected as feature points by the feature location method, and the x coordinate and y coordinate of the nostrils are used to indicate the directions of head motions to determine whether a head nod or a head shake occurs. Both nostril detection and head movement detection obtain ideal recognition results. We further quantify head movement into fuzzy input by utilizing the frequency of head movement.

Chapter 5 illustrates eye gaze detectors that can analyze the subject's gaze direction. The main objective is to quantify subject’s gaze status into numerical fuzzy input. Our approach is knowledge based, and preliminary results are taken as assumptions. After we detect the locations of the nostrils and the two pupils, the relations between these four points are studied and evaluated into a parameter in order to represent the gaze status.

Chapter 6 introduces the facial expression recognition method using the multi-step integral projection for image segmentation and Gabor transformation for feature detection. A support vector machine (SVM) learning algorithm is applied to classify facial expressions in six categories.

Finally, Chapter 7 concludes the thesis by summarizing the contributions of this work as well as the possible improvements for future work.

Chapter 2

Literature Review

2.1 Introduction

2.1.1 Approaches for Categorizing Emotions

Emotion is a complex psychophysiological experience that results in physical and psychological changes that influence an individual's behaviour [1]. It is caused by an external (environment) stimulus that leads to an internal (physiological) reaction [1]. Based on research in psychology, emotion classification can be divided into three major approaches: categorical approach, dimensional approach, and appraisal approach [6].

- **Categorical approach**

This approach is based on the research on emotion theory from Darwin [7], interpreted by Tomkins [8] and then supported by Ekman [9]. According to this approach, there exist a small number of emotions that can be universally recognized. Ekman [10] applied various experiments on human judgment of still images of deliberately expressed facial expressions and concluded there are six basic emotions that can be recognized

universally. These six emotions are happiness, sadness, surprise, fear, anger, and disgust. Although some researchers have suggested a different number of basic emotions, ranging from 2 to 18 [11], [12], there has been considerable agreement on the six basic emotions. To date, Ekman's theory on universal nonverbal emotional expression of the six basic categories has been the most commonly adopted approach in research on automatic emotion detection.

However, some researchers have argued that it is necessary to go beyond the six discrete emotions [11], [51], [78]. Moreover, some researchers, for example, Baron-Cohen and his colleagues [13] have explored cognitive states (e.g., agreement, disagreement, thinking, concentrating, and interest) and their use in daily life. They analyzed multiple asynchronous information sources such as facial actions and head movements. They reported that cognitive states occur more often in human day-to-day interactions than the six basic emotions [13]. The disadvantage of the categorical method is that each emotional display is classified into a single category, therefore complex emotional states or secondary emotions will be too difficult to determine [14].

- **Dimensional approach**

According to the dimensional approach, emotional states are not independent from each other. On the contrary, they are related to one another in a systematic manner [15]. In this approach, emotional variability is described by three dimensions: valence, arousal, and power [15], [16]. The valence dimension stands for positive or negative degree of the emotion, and it ranges from unpleasant feelings to pleasant feelings. The arousal dimension stands for how excited or apathetic the emotion is and it ranges from sleepiness or boredom to frantic excitement. The power dimension stands for the degree of power or sense of control over the emotion. In 1980, Russell [17] proposed the Circumplex of Affect using a bipolar circular configuration. They stated that each basic emotion corresponds to a bipolar unit as a part of the emotional continuum. The proposed polars are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant). Figure 2.1 illustrates the proposed emotional space consisting of four quadrants: low arousal positive, high arousal positive, low arousal negative, and high arousal negative. According to the differences of the valence and arousal, emotions can be distinguished and plotted on the two dimensional place.

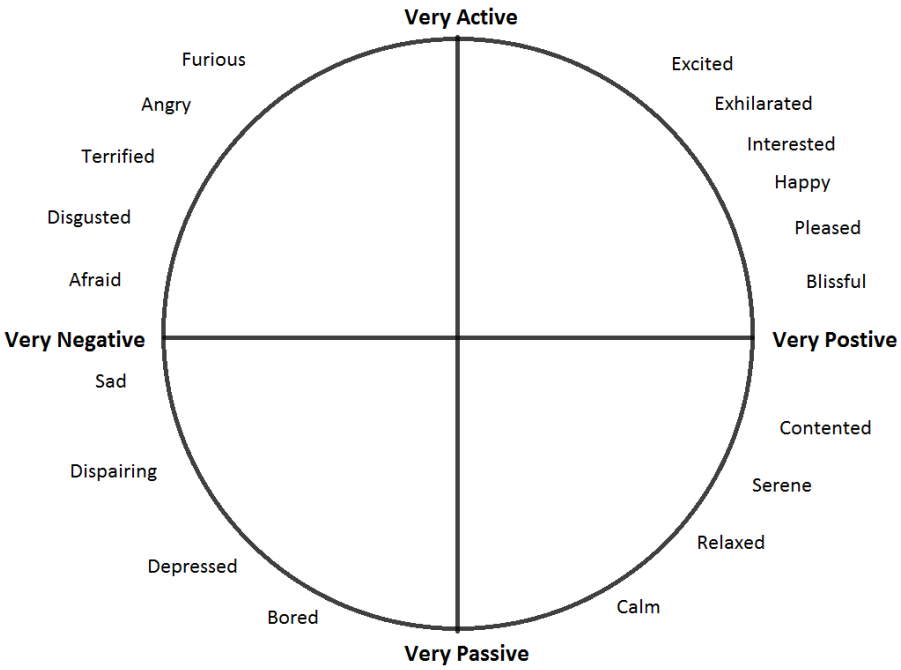


Figure 2.1 Distribution of emotions on dimensional space

The advantage of the dimensional approach is that observers can designate their intuition of each stimulus on several continuous scales. However, the usefulness of the dimensional approach also has been challenged by discrete emotion theorists who have argued that the reduction of emotion space to two or three dimensions is reductive and results in loss of information [8], [10]. In addition, although some of the basic facial expressions proposed by Ekman, such as happiness or sadness, appear to easily distinguish in the dimensional space, some others such as fear and anger become not easy to distinguish and some facial expressions may lie outside the dimensional space (e.g., surprise). Moreover, it is also unclear how to distinguish the position of some cognitive states such as confusion on the space.

- **Appraisal approach**

Appraisal approach proposed by Scherer [18] can be seen as an extension of the dimensional approach. This approach focuses on identifying emotion according the interpretations of events of the outside world that cause emotions, rather than events themselves [18]. This approach investigates emotions through changes in all relevant aspects such as cognition, physiological reactions, motivation, feelings, and so on.

The advantage of appraisal approach is that they do not limit emotions to a defined number of discrete categories or to a few fundamental dimensions. As an alternative, they

focus on the changeability of emotional states, as produced by various types of appraisal patterns. Emotion is described through a set of stimulus evaluation checks, including the novelty, compatibility with standards, intrinsic pleasantness, and goal-based significance [18]. Therefore, differentiating between various emotions and modeling individual differences become possible. Because this method requires complex and sophisticated measurements of change, how to use the appraisal approach for automatic emotion detection in HCI remains an open research question [18].

Despite over a century of research, all of the issues mentioned above, particularly the issue of which psychological model of emotion is more appropriate for which context, are still under debate [19], [6].

2.1.2 Emotion Signals

Detecting and interpreting human emotion is not an easy task for computers, since emotion itself is complex, and the range of verbal and nonverbal emotional cues is large. Human behaviour and emotion can be inferred and interpreted from the conveyed information through many modalities. We summarized a hierarchical relationship of human emotional cues as shown in Figure 2.2. In this figure, the hierarchical structure among the signals is displayed. For example, human behaviour cues consist of human emotional cues and other behavioural signals such as yawning, grunting, sniffing, etc. Human emotional signals include speech information, text information, and other non-verbal cues. The detailed lists below show the relationship and the content of these signals:

- Human Behavioural Signals
  - Yawning, grunting, sniffing, coughing, etc.
  - Human emotional signals
- Human Emotional Signals
  - Speech (explicit linguistic messages)
  - Text
  - Non-verbal signals
- Non-verbal Signals

- Physiological signals (e.g., brain activity, heart rate, blood flow, etc.)
- Speech (implicit linguistic messages, e.g. pitch and inflection)
- Sound (e.g., laughing, crying, etc.)
- Visual signals

➢ Visual Signals

- Tears, flush, etc.
- Body languages

➢ Body Languages

- Hand gestures
- Body postures
- Body language of the head

➢ Body Language of the Head

- Facial expressions
- Head movements
- Eye gaze

2.2 Methods of Emotion Recognition

Emotional information can be conveyed by a wide range of multimodal emotion signals as presented in the previous subsection. In emotion detection for HCI, all of the emotion signals can be categorized into four types, which are shown in Figure 2.3.

In current emotion detection systems, researchers are attempting to detect emotion mainly from the following modalities: facial expression, audio (voice), body posture, physiological information, text, and multimodal cues. In this thesis, the review of the current body of literature of emotion recognition systems is organized with respect to the above modalities. Each modality has its own advantages and disadvantages based on its use as a viable emotion recognition channel.

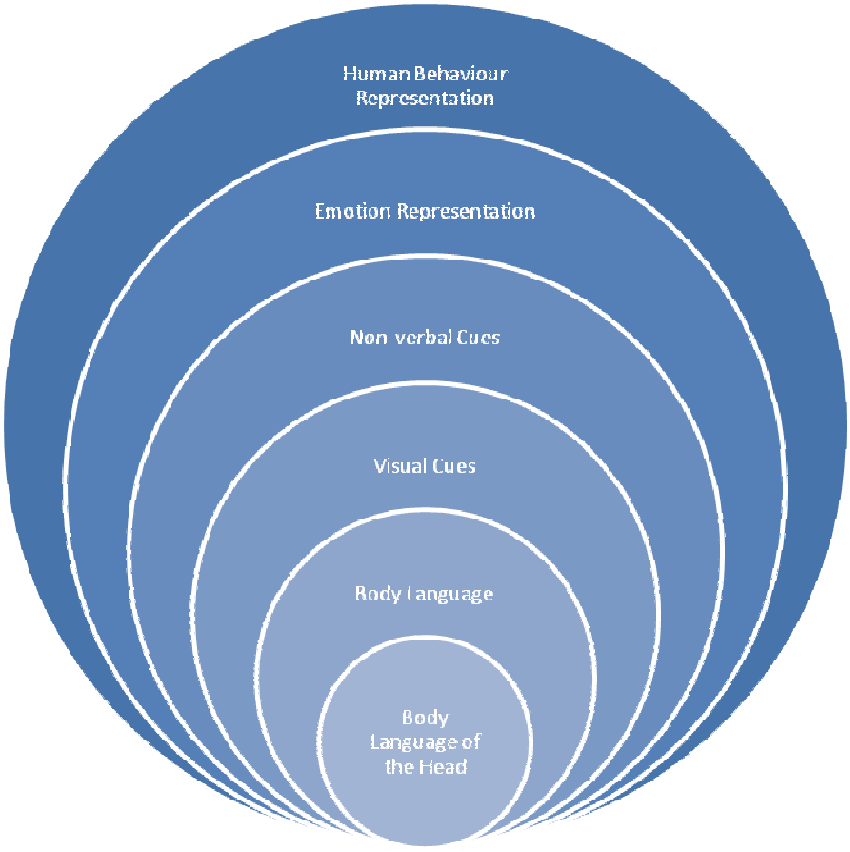


Figure 2.2 The hierarchical relationship of human emotional cues

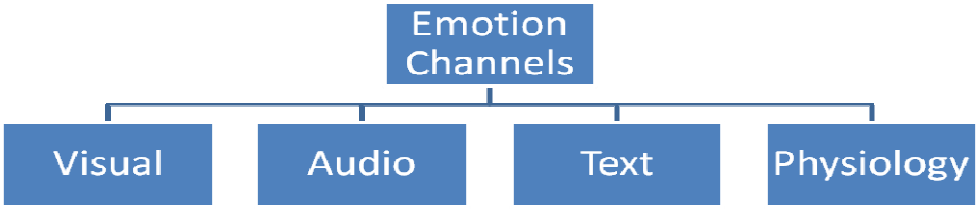


Figure 2.3 Four types of channels that can convey emotion

2.2.1 Emotion Recognition from Facial Expressions

The attempt to recognize emotion from facial expressions is based on the theory that there is a distinctive facial expression associated with each basic emotion [9], [20], [21]. The expression is triggered for a short duration when the emotion occurs. Therefore, detecting an emotion is simply a matter of detecting its prototypical facial expression. Based on this theory, the mainstream of researchers on machine analysis of human emotion recognition has focused on facial expression recognition, specifically in

recognizing the six basic emotion categories: happiness, sadness, fear, anger, surprise, and disgust [22], [23].

The human face contains major communicative cues of human emotions. It can not only show facial appearance such as age and gender, but also can infer the affective states and intentions of the person you are communicating with. There are four kinds of signals that the human face can convey [165]:

- Static facial signals: These signals represent relatively permanent facial features of the face (e.g., the bone structure). These signals contribute to the individual appearance of a person and can be used for personal identification recognition.
- Slow facial signals: These signals contribute most in age detection since they characterize the transforms in the face's appearance that happen slowly over time (e.g., the development of permanent wrinkles).
- Artificial signals: These signals are exogenous aspects of the face that can supply supplementary information that contributes most to gender recognition (e.g., glasses and cosmetics).
- Rapid facial signals: These signals correspond to temporal transforms in neuromuscular activity that results in optically noticeable changes in facial appearance (e.g., facial expressions, head nods, and eye gaze). These signals contribute most to facial expression recognition and emotion detection.

The most communicative manner human express their emotional state is from facial expressions. Research in social psychology [24], [21] has suggested that facial expressions form the major modality in human communication and are a visible manifestation of the affective state of a person. Since facial expressions are our straight, natural, and most excellent way of expressing emotions, machine learning of facial expressions forms an indispensable part of human-computer interface designs [25], [26].

There are two major approaches in the present research on automatic analysis of facial expressions:

- Facial affect detection.
- Facial muscle action (action unit) detection.

These two streams come from two major approaches of facial expression analysis methods in psychological research [27]:

- Message judgment approach: aims to infer what underlies a displayed facial expression.
- Sign judgment approach: aims to describe the surface of the shown behaviour, such as facial component shape.

Researchers agree, for the most part, that most facial expressions are learned like language and have culturally specific meanings that rely on context for proper interpretation [28]. On the other hand, there are a limited number of distinct facial expressions that appear to share similar understandings across all cultures [10], [22]. Therefore, these universally displayed facial expressions are widely used in emotion detection approaches, which are as follows: happiness, sadness, anger, fear, disgust, and surprise [10], [27].

Sign judgment approaches use the Facial Coding System (FACS) [29], [30] to manually label facial actions. FACS describes the facial expression changes with actions of the muscles that produce them. It describes 44 facial action units (AUs) that are considered the smallest noticeable and detectable facial movements (see Figure 2.4). It also provides the rules for recognition of AU's temporal divisions (onset, apex and offset) in a facial expression video. By applying FACS, almost all anatomically possible facial displays can be manually coded by decomposing it into AUs.

So far, Ekman's theory of the six basic facial expression and the FACS are the most commonly used methods in vision-based facial expression recognition systems designed to analyze human affective behaviour [25], [23].

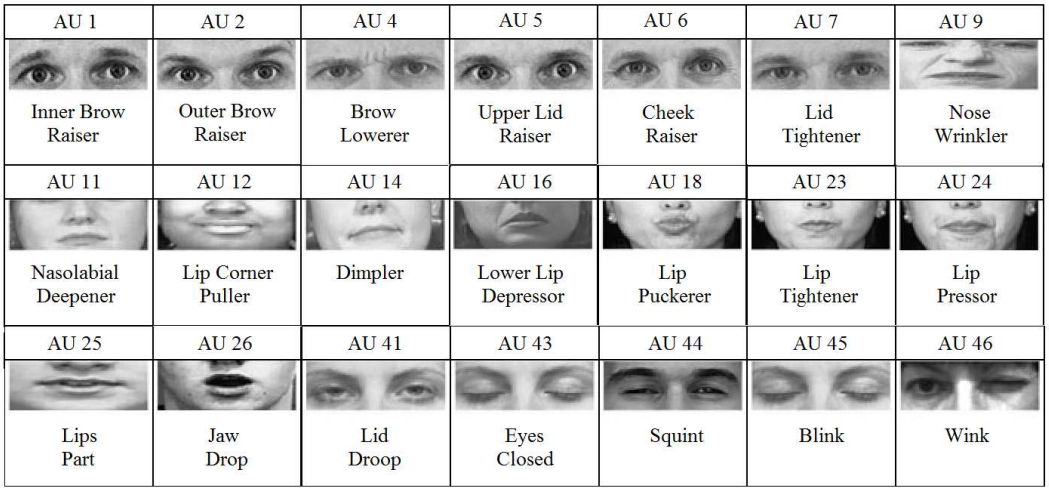


Figure 2.4 Examples of facial action units (AUs) from Cohn and Kanada database

A thorough review of the state of the art of the vision-based facial expression methods can be found in Zeng et al.'s study [23]. Recently, there has been a shift toward recognition of deliberate and exaggerated facial expression displays to the analysis of spontaneous facial expression [32], [33], [34], [35], [36], [37], [38], [39], [40]. This shift in facial expression recognition is aimed toward subtle, continuous, and natural interpretations of emotional displays recorded in real-world environments. In particular, some of them are analyzing the recognition of AUs instead of emotions from spontaneous facial displays [32], [41], [37], [38]. However, although there has been remarkable progress in this area, the reliability of current automatic AUs recognition systems does not match humans [42], [43], [44]. On the whole, recognizing emotion from facial expressions has the following advantages and disadvantages.

Advantages:

- Most natural way to identify emotional states.
- Available databases.
- Low cost and intrusiveness of the user.

Disadvantages:

- Cannot provide context information thus sometimes results are misleading.
- Detection results dependent on image or video quality.

2.2.2 Emotion Recognition from Voice

Recognizing emotion from vocal information is also influenced by the basic emotion theory. Speech conveys emotional information through two types of messages [45]:

- Explicit (linguistic) messages.
- Implicit (paralinguistic) messages.

Explicit messages are the transcriptions of oral communication, and implicit messages are features that reflect the way the words are spoken. If we consider the verbal part (*what* is said) only, without considering the manner in which it was spoken (*how* it is said), we might miss important information of the pertinent utterance and even possibly misunderstand the spoken message [45].

The speaker's emotional state can be inferred directly from the transcriptions of words that were summarized in some emotional word dictionaries and lexical affinity [46]. However, findings show that spoken messages are rather unreliable in analyzing human emotional states [47]. First, predicting a person's word option and the related objective is difficult. Even in highly constrained situations, different people choose different words to express exactly the same thing. Moreover, the relationship between linguistic content and emotion is language dependent and generalizing from one language to another is difficult to achieve [11].

Implicit messages can also convey emotional information. Research in psychology and psycholinguistics provides numerous results on acoustic and prosodic features that can be used to infer the basic emotional states of a speaker [48]. The most reliable finding is that pitch appears to be an index into arousal [45]. A number of works has been focusing on mapping audio expression to dimensional models. Cowie and his colleagues applied valence activation space to analysis and indicate emotions from speech [49], [50]. Scherer et al. have also indicated how to analyze affective states on vocal information using the appraisal approach [6], [19]. A comprehensive discussion of the literature review of audio-based emotion detection is summarized by [23]. However, the interpreting of paralinguistic signals has not yet been fully understood [48]. Although listeners seem to be able to accurately recognize some basic emotions from prosody and some behavioural states such as depression, anxiety, boredom, and interest from non-verbal vocalizations such as laughs, cries, and yawns, researchers have not provided an optimal set of voice signals that can discriminate emotions reliably [48]. There is still some ambiguity regarding how different acoustic features communicate the different emotions. In addition, the detection accuracy rates from speech are somewhat lower than the detection accuracy rates from facial expressions for the six basic emotions [45]. For example, although sadness, anger and fear are best recognized by voice, disgust shows a low recognition rate. The following are the advantages and disadvantages of recognizing emotions using audio information:

Advantages:

- Easy to achieve.
- Low cost and non-intrusiveness.
- Available databases.

- More apt to meet the needs of real-world applications compared to vision-based approaches.

Disadvantages:

- Greatly affected by different types of languages.
- Lower detection accuracy compared to facial expression.
- The implicit messages have not yet been fully understood.

2.2.3 Emotion Recognition from Body Posture

Darwin [7] was the first to propose several principles describing in detail how body language and posture are associated with emotions in humans and animals. Although Darwin's research on emotion analysis heavily focused on body language and posture [7], [52], [53], state of the art emotion detection systems have overlooked the importance of body posture compared to facial expressions and voice recognition and the expressive information carried by body posture has not yet been adequately explored [54].

The human body is relatively large and has multiple degrees of freedom. It can provide humans with the capability of assuming a myriad of unique configurations [55]. These static positions can be simultaneously merged into a large amount of movements that let body posture a prospectively ideal emotion communication channel [56], [57]. Moreover, body posture can provide information that sometimes cannot be transmitted by conventional nonverbal channels such as facial expressions and speech. For instance, the emotional type of a person can be distinguished from a long distance by displaying body posture, whereas detection from other channels of characters is difficult or unreliable [58]. The greatest advantage of recognizing emotion from posture information is that the gross body expression is unaware, unintended and, therefore, not susceptible to the surrounding environment at least compared to facial expressions and speech [56], [57], [58]. Ekman and Friesen [4] in their study of deception, show that it is much easier to disguise deceit through facial expression than the less controllable body posture channel.

In 2004, Coulson produced the six basic emotions from descriptions of postural expressions of emotion in order to investigate the attribution of emotion to static body postures by using computer-generated figures [56]. His experimental results show that, in general, the recognition of human emotion from body posture is as good as the

recognition from voice information, and some postures are distinguished as successfully as facial expressions.

Van et al. [59] conducted a study investigating emotional body postures of the six basic emotions and how they are perceived. Experimental results show good recognition of emotions where anger and fear body expressions are less accurately recognized compared to some emotions in bodily expressions (e.g., sadness).

Body posture can communicate discrete emotion categories as well as emotional dimensions [56], [57]. Therefore, another stream of emotion recognition from body posture is focusing on mapping body expression into emotion dimensional space. Kleinsmith et al. [60] found that emotional dimensions including scaling, arousal, valence, and action tendency, were used by human observers to distinguish among postures. They reported that low level posture signals such as orientation (e.g., orientation of shoulder axis) and distance (e.g., distance between left elbow and left shoulder) could effectively distinguish between the emotional dimensions.

Mota and Picard [61] presented a system to automatically analyze naturally displayed postures to infer the behavioural states of a user in a learning environment. They collected body postures by using two matrices of pressure sensors mounted on the seat and back of a chair. A neural network was applied to classify nine static postures, such as leaning back, sitting upright, etc. The system then analyzed temporal transitions of posture patterns to estimate the interest level of children (high interest, low interest, and taking a break).

D'Mello and Graesser [62] extended Mota and Picard's work by developing an intelligent tutoring system to detect cognitional and emotional states from the students' gross body movements. The system measures the body pressure of the learner on the seat and back of the chair. They extracted two sets of features obtained from the pressure maps. The first class is the average pressure applied with the magnitude and direction of transforms in the pressure during the occurrence of affection. The second class examined the spatial and temporal properties of naturally occurring pockets of pressure. Machine analysis methods were used to detect emotions including boredom, confusion, delight, flow, and frustration.

In general, detecting emotional information through body posture and movement is still a relatively unexplored and unresolved area in psychology [56], [57]. Further research is needed to obtain better insight on how they contribute to the perception and

recognition of various emotional states for HCI. The major pros and cons of posture-based emotion detection are the following:

Advantages:

- Can be distinguished from long distances.
- Gross body motions are ordinarily unconscious.

Disadvantages:

- Relatively unexplored and unresolved area in psychology.
- Expensive and intrusive equipment needed.

2.2.4 Emotion Recognition from Physiology

Numerous findings in psychophysiology suggest that the activation of the autonomic nervous system changes when emotions are elicited [63]. While the visual channel such as facial expressions and body postures provide visible proof of emotional arousal, bio-signals provide an invisible proof of emotional arousal [64]. The signals are usually referred to as physiological or bio-signals [64], [65] are used in the emotion sensing research field to identify emotions can be listed and described as follows.

- Electroencephalography (EEG) measures brain activity that can provide an invisible proof of emotional arousal.
- Galvanic Skin Response (GSR) provides a measurement of the of skin conductance (SC). SC increases linearly with a person's level of overall arousal or stress.
- Electromyography (EMG) measures the muscle activity or frequency of muscle tension, and has been shown to correlate with negatively valence emotions.
- Blood Volume Pulse (BVP) is an indicator of blood flow. Since each heart beat (or pulse) presses blood through the vessels, BVP can also be used to calculate heart rate and inter-beat intervals. Heart rate increases with negatively valence emotions, such as anxiety or fear.
- Skin temperature (ST) describes the temperature as measured on the surface of the skin.
- Electrocardiogram (EKG or ECG) signal measures contractile activity of the heart. This can be recorded either directly on the surface of the chest or alternatively on

the limbs. It can be used to measure heart rate and inter-beat intervals to determine the heart rate variability (HRV). A low HRV can indicate a state of relaxation, whereas an increased HRV can indicate a potential state of mental stress or frustration.

- Respiration rate (RR) measures how deep and fast a person is breathing. Slow and deep breathing indicates a relaxed resting state while irregular rhythm, quick variations, and cessation of respiration corresponds to more aroused emotions like anger or fear.

Several HCI applications focus on detecting emotional states by using machine learning techniques to identify patterns in physiological activity. Table 2.1 summarizes some recent work on applying physiological signals for emotion detection. Researchers stated that physiological or bio-signals offer great changes for automatic emotion detection [66]. However, developing their full prospective has been unfeasible to date because of the lack of consensus among psychologists about the nature, theories, models, and specificity of physiological models for each emotion-space dimension. The high cost, time resolution, and complexity of setting up experimental protocols are still issues that hinder the development of practical applications of this technique. In the future, establishing standardization on key areas such as stimulus for the identification of physiological patterns, physiological measures, features to analyze, and the emotional model to be used will greatly advance the state-of-the art in this field [66].

Advantages:

- Can provide invisible cues that cannot be concealed.

Disadvantages:

- Lack of consensus of physiological patterns (neural models of emotion).
- Expensive and intrusive sensors needed.
- Noise in the signals.

Table 2.1 Recent studies of emotion detection from physiology signals

Authors/Year	Modality	Classifier	Results	Stimulus
Calvo et al. 2009 [67]	EKG, SC, EMG	SVM	8 categories	Self elicited
Alzoubi et al. 2009 [68]	EEG	SVM, NB, KNN	10 categories	Self elicited
Bailenson et al. 2008 [69]	ECG, SC, Face	SVM	2 categories	3rd & Films
Liu et al. 2008 [70]	ECG, SC, EMG, ST	SVM	3 categories	1st and 3rd person
Heraz and Frasson 2007 [71]	EEG	Nearest Neighbours, Bagging	3 dimensions	IAPS
Villon and Lisetti 2006 [163]	HR, SC	Not available	Dimensional model	No subjects evaluated
Wagner et al. 2005 [72]	EKG, SC, EMG	PCA	4 categories	Self-selected songs
Kim et al. 2004 [73]	EKG, ST, SC, HRV, RR	SVM	4 categories	Audio-visual, self-evaluation
Nasoz et al. 2004 [74]	SC, HR, ST	KNN, DFA, MBP	6 categories	Movie clips selected by panel
Haag et al. 2004 [75]	EKG, EMG, SC, ST, BVP	MLP	Dimensional model	IAPS

Classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Principal Component Analysis (PCA), Discriminant Function Analysis (DFA), Marquardt Backpropagation (MBP), Multilayer Perception (MLP).

2.2.5 Emotion Recognition from Text

Emotion recognition from text refers to written language and transcriptions of oral communication. Psychologists and anthropologists seek to discover associations on how individuals from different cultures communicate [76], [77].

Osgood proposed three dimensions according to text information: evaluation, potency, and activity [76]. Evaluation refers to how a word refers to an event (e.g., pleasant or unpleasant); potency quantifies how a word is associated to an intensity level (e.g., strong or weak); activity refers to whether a word is active or passive. These dimensions are similar to the dimensions of emotion space (valence and arousal), which are considered the fundamental dimensions of emotion experience [78], [79].

Another approach involves a lexical analysis of text to identify words that are predictive of the emotional states of the user [80], [81], [82], [83], [84], [85]. Some of these approaches depend on the Linguistic Inquiry and Word Count (LIWC) [86], a computer tool that can analyze bodies of text by applying dictionary-based

categorization.

Some researchers assume that people using the same language would have a similar conception for different discrete emotions. For instance, Wordnet is a lexical database that consists of English terms and is widely used in computational linguistics research [87]. Table 2.2 lists some works on text analysis for emotion detection.

Advantages:

- Easy to achieve.

Disadvantages:

- May lead to wrong results without the context of paralinguistic messages.
- Not a very natural way for human emotion recognition comparing to other modalities.

Table 2.2 Text analysis for emotion detection

Authors/Year	Evaluation	Classifier	Results	Text
D'Mello et al. 2009 [88]	1st and 3rd	Cohesion Indices	4categories	Dialogue logs
Danisman et al. 2008 [89]	1st	SVM, NB	5 categories	ISEAR questions
Strapparava et al. 2008 [90]	3rd	Latent Semantic Analysis (LSA)	6 categories	News stories
Lin et al. 2006 [91]	3rd	NB, SVM	2 categories	Political articles
Alm et al. 2005 [92]	1st	Winnow Linear	7 categories	Children stories

Classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), Latent Semantic Analysis (LSA).

2.2.6 Emotion Recognition from Multimodality Approaches

Humans express emotions through simultaneous combinations of verbal and nonverbal acts including facial expressions, head or eye movements, hand signals, body postures, and so forth. Attempting to infer emotion from facial expressions alone does not always lead to accurate conclusions. Although researchers commonly advocate combining information from different modalities, few systems have been implemented [93], [23], [54]. Nevertheless, the advantage of applying multiple modalities of emotion detection has been recognized as the next step for the mostly single modality approaches.

2.2.6.1 Data Fusion

In the process of emotion detection, fusion refers to combining and integrating all incoming single modalities into one representation of the emotion expressed by the user. When it comes to integrating the multiple modalities the major issues are [94]:

- When to integrate the modalities (at what abstraction level to do the fusion)
- How to integrate the modalities (which criteria to use)

According to the above factors, there are two typical data fusion methods: feature level fusion and decision level fusion [94].

- **Feature level fusion**

Feature level data fusion is performed on the set of features extracted from each modality. It is appropriate for closely coupled and synchronized modalities (e.g., speech and lip movements) [94]. Therefore, feature level data fusion appears not to normalize well when the modalities substantially differ in temporal characteristics (e.g., speech and posture). Therefore, the features extracted from feature level should be synchronous and compatible.

- **Decision level fusion**

Decision level data fusion is based on the assumption that different modalities are mutually independent. In this approach, each modality is classified separately and the outputs of each classifier are integrated at a later stage in order to obtain a global view across the expressed emotional behaviour (see Figure 2.5). The decision level data fusion is the most commonly applied approach for multimodal HCI, especially when modalities differ in temporal characteristics [164], [25].

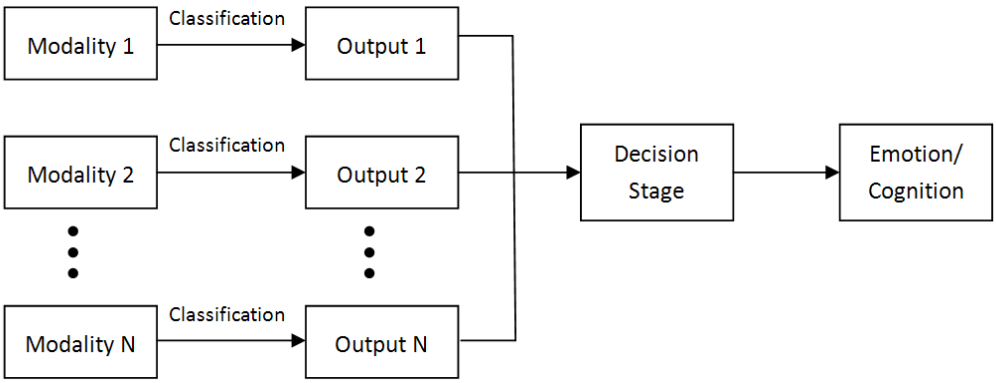


Figure 2.5 The process of decision level fusion

The typical reasons to use decision level fusion (i.e., late integration) instead of feature-level fusion (i.e., early integration) can be summarized as follows [31]:

- ✓ The feature concatenation used in feature level fusion results in a high dimensional data space, causing large multimodal dataset.
- ✓ Allows asynchronous processing of the available modalities.
- ✓ Provides greater flexibility in modeling, e.g., possible to train different classifiers on different data sets and integrate them without retraining.
- ✓ Recognizers can be utilized for single modalities.
- ✓ More adaptable and extendable, e.g., new modalities can be added based on context information.

2.2.6.2 Selected Works on Multimodal Approaches

Scherer and Ellgring [95] analyzed emotion portrayals of combining facial expression, vocal features, gestures, and body movements to discriminate among 14 emotions. The results showed that the combined system obtained a much higher recognition rate than a single channel. However, when only 10 features are selected using the combined model, the classification accuracy dropped to only 54%. Therefore, it is difficult to say whether the combined channels led to enhanced classification rates compared to single channels because the number of features between the single channel and combined channel was not equivalent.

Castellano et al. [96] presented a multimodal approach of detecting eight emotions, including some basic emotions including irritation, despair, etc., using the combined information of body movement, facial expressions, speech, and gestures. The multimodal

classifier resulted in an improvement of 17% in accuracy compared to the best single-channel system.

Kapoor and Picard [3] proposed a multimodal, contextually grounded, probabilistic system to recognize interest or disinterest of children in solving a puzzle. The system utilized the upper and lower facial features and the body posture features as well as some contextual information. The combined information achieved a recognition rate over 86%, which is higher than the single channel facial features. However, it has a similar recognition rate compared to body posture alone, indicating that the other channels are redundant with the body posture channel.

In 2007, Kapoor et al. [98] extended their system by adding other modalities including a skin conductance sensor and a pressure-sensitive mouse. It predicts children's emotion of frustration when interacting with an avatar playing the Hanoi problem. This new approach was tested on 24 participants and obtained an accuracy of 79%. Although this study illustrated that combining multiple nonverbal channels can provide important information for detecting emotional states, it did not report the single channel detection accuracy. Therefore, it is difficult to assess the specific advantages of combining multiple channels.

Arroyo et al. [97] considered the combination of various physiological sensors (seat pressure sensor, galvanic skin conductance sensor, pressure mouse sensor) as well as facial features and context information to infer the student's emotions (levels of confidence, frustration, excitement, and interest). This system reported that facial feature and context information explained 52, 29, and 69% of the variance of confidence, interest, and excitement, respectively. The seat pressure sensor and context contribute most in frustration recognition. Their research concluded that monitoring facial and context feature provide the best models, while the other modalities did not provide additional advantages.

More recently, D'Mello and Grasesser [99] developed a multimodal system that combined conversational cues, gross body language, and facial features. The recognition results showed that the accuracy of multimodal channels was statistically higher than the best recognized single-channel model for the fixed emotion expressions (but not for spontaneous emotion expressions). Although the combined channels yielded improvement in recognizing some emotions, they may provide redundant and additive information for others. Table 2.3 summarizes some selected typical works using

multimodal channels for emotion detection. In summary, although detecting emotional states through multiple modalities can provide more information than a single channel, there still remain problems to overcome. The following lists the pros and cons for multichannel-based approach.

Advantages:

- Can infer complex emotional states.

Disadvantages:

- Multidisciplinary knowledge required.
- How to handle conflict information conveyed by modalities is still an open question.
- Difficult in optimizing information with high disparity in accuracy.

Table 2.3 Recent works for emotion detect from multimodal approach

Authors/Year	Modality	Fusion	Classifier	Results	Stimulus
D'Mello and Graesser, 2010 [99]	Conversational cues, body language, and facial features	Feature level fusion	Linear Discriminant Analyses (LDA)	Boredom, engagement, confusion, frustration, delight, and neutral	Intelligent tutoring system
Arroyo et al. 2009 [97]	facial features, physiological signals (seat pressure, galvanic skin conductance, pressure mouse)	Decision level fusion	Support Vector Machines, Gaussian Process Classification	Confident, frustrated, excited, interested	Real public schools educational settings, data collected by sensors (compare to self-report results)
Caridakis et al. 2008 [100]	Facial expression, body gestures, and audio	Decision level fusion	Neural Network	Neutral and four A-V quadrants	TV programs
Castellano et al. 2008 [96]	Facial expression, body movement, gestures, and speech	Feature level fusion and decision level fusion	Bayesian classifier	Anger, despair, interest, irritation, joy, pleasure, pride, and sadness	Actors (self elicitation)
Kim 2007 [101]	Speech and physiological signals (EMG, SC, ECG, BVP, RSP, and Temp)	Integrating results from feature level fusion and decision level fusion	Modality-specific LDA-based classification; a hybrid fusion scheme where the output of feature level fusion is fed as an auxiliary input to the decision level fusion	Either of the four A-V quadrants	Quiz shows on TV
Kapoor et al. 2007 [98]	Facial features, head gestures,	Decision level fusion	Support Vector Machines, Gaussian	Frustration	Solving a puzzle on the

	physiological signals (seat pressure sensor, galvanic skin conductance sensor, pressure mouse sensor)		Process Classification		computer
Karpouzis et al. 2007 [102]	Facial expression, hand gestures, vocal features	Decision level fusion	Neural Network	Negative vs. positive, active vs. passive	TV programs
Kulic et al. 2007 [103]	Tactile, physiological (heart rate, perspiration rate, and facial muscle contraction)	Not reported	Hidden Markov Model	Six emotion categories (low/medium/high-valence/ arousal)	Robot motions
Kapoor and Picard, 2005 [3]	Facial features, head gestures, postures	Decision level fusion	Support Vector Machines, Gaussian Process Classification	Level of interest	Solving a puzzle on the computer
Kaliouby and Robinson, 2005 [104]	Head movement, and facial features	Decision level fusion	Hidden Markov Models, Dynamic Bayesian Networks	Agreeing, concentrating, disagreeing, interested, thinking, and unsure	Actors (Posed)

2.3 Human Behaviour Recognition from Body Language of the Head

Kaliouby and Robinson [132] examined the use of the combined information from facial expressions and head movements to infer complex cognitive states from the Mind Reading DVD [133]. The Mind Reading DVD is a computer-based guide that covers a range of emotion and cognition states. It is developed by a team of psychologists who aimed to help individuals diagnosed with autism spectrum. It records 412 mental state concepts that that are divided into 24 classes. Their work focused on the six classes (agreeing, concentrating, disagreeing, interested, thinking and unsure) out of the 24 classes for the emotion recognition.

Some works have been found by utilizing body language of the head to detect human behavioural states such as fatigue, attentive, and non-attentive. Ji et al. [134] employed machine learning methods to detect human behaviour of drowsiness during driving by incorporating information from visual information including facial expressions (expressionless face or yawning), eyelid movement (blink frequency), line of sight (eye openness or closure), head motions (head tilts, head down or sideway), as well as contextual information (sleep time, workload, physical condition, etc.).

Vural et al. [162] presented a system for automatic detection of driver drowsiness for spontaneous behaviour during real drowsiness episodes, where participants played a driving video game using a steering wheel. They integrated blinking, yawing, eye openness, head tilts, eye brow raise, as well as other facial movements to detect fatigue.

Asteriadis et al. [135] presented a system that can detect the behavioural states in the context of children reading an electronic document. The movements of the head, eye, and hand were tracked to estimate the children's level of interest. They applied the neuro-fuzzy techniques to classify the behavioural states into attentive or non-attentive states during reading.

Ba et al. [161] attempted to recognize the visual focus of attention of participants by analyzing head pose (pan and tilt), and gaze direction (utilized the head orientation to estimate gaze instead of analyzing pupil direction) in general meeting scenarios. Their work focused on remote distance attention analysis (by head pose tracking) when high resolution close-up views of the eyes cannot be obtained.

Despite the success of the existing works, none of the previous research simultaneously considered all modalities from the body language of the head to infer human emotions. The goal of this thesis is to explore new solutions for HCII by integrating information from the body language of the head to infer emotional and cognitive states. In particular, we concentrated on the combination of facial expression, eye gaze and head movement by applying soft computing techniques [136], [137], [138]. A two-stage approach is proposed. The first stage analyzes the explicit information from the modalities of facial expression, head movement, and eye gaze separately. In the second stage, all these information are fused to infer the implicit secondary emotional states. To analyze head movements, not only the direction but also the frequency of head movement is observed. Eye gaze direction is also integrated with other head information to analyze emotional states.

2.4 Emotion Detection applying Fuzzy Techniques

2.4.1 Introduction to Fuzzy Techniques

Fuzzy logic can be seen as a generalized classical logic. Classical logic (two-valued logic) concerns propositions that are either true or false. The truth value set of classical

logic has two elements: 0 representing false and 1 representing true. In 1920, Lukasiewicz introduced many-valued logic where the truth value set has more than two elements besides 0 and 1 [158]. By using fuzzy sets and fuzzy relations in the system of many-valued logic, fuzzy logic is derived from many-valued logic. A methodology is given by fuzzy logic for treating linguistic variables and expressing modifiers like very, fairly, not and so on.

Fuzzy logic makes reasoning with imprecise and vague propositions which cope with the natural language easily. It reflects both the rightness and vagueness of natural language in common-sense reasoning. Linguistic variables are variables which take values of words or sentences in natural or artificial languages [137]. For example, age is a word in natural language. Let age be a linguistic variable taking values from a set of words: very young, young, middle age, old, very old. These values are called terms of linguistic variable age and described by fuzzy sets with corresponded membership functions on a universal set.

A fuzzy inference system (FIS) is a system that uses fuzzy logic to make decisions or solves problems in a particular field by using knowledge and analytical rules [136]. It is developed for reasoning based on a collection of membership functions and inference fuzzy rules.

The triangular and trapezoidal membership functions are most popular choices. The mathematical description of a general trapezoidal membership function is given by

$$\mu(x) = \begin{cases} 1, & x_2 \le x \le x_3 \\ (x - x_1)/(x_2 - x_1), & x_1 \le x \le x_2 \\ (x_4 - x)/(x_4 - x_3), & x_3 \le x \le x_4 \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

where x_1, x_2, x_3 and x_4 are the trapezoidal membership function parameters, x is quantitative input and $\mu(x)$ is the membership function.

Inference rules are a set of “if … and ... then” rules. In FIS, inference rules stem from the knowledge of human experts, the preference of clients, or common sense of everyday life. They can be redesigned at any time when knowledge base changes.

2.4.2 Brief Review of Emotion Detection using Soft Computing Techniques

Soft computing techniques are widely used in the area of pattern recognition; for example, in [139] fuzzy classifier was applied for face recognition. In [140], [141], [142], fuzzy techniques were used in biometric applications.

With the development of affective computing [5] and HCII, the fuzzy logic technique has been used for emotion recognition. Developing linguistic rules is the most important step in developing a fuzzy logic system. These rules can be knowledge-based, which are provided by experts [143], [136]. When expert knowledge is not sufficient, the rules can also be extracted from a given training data set (data-driven-based) by using computer learning algorithms [143], [136]. Chakraborty and Konar [144] used a knowledge-based fuzzy system to recognize emotion from facial expressions. By dividing face images into localized regions, facial features including eye opening, mouth opening, and the length of eyebrow constriction were extracted, fuzzified, and mapped into facial expressions. Contreras et al. [145] presented a knowledge-based fuzzy reasoning system that could recognize the intensity of facial expressions by using the facial action units and facial animation parameters. Esau et al. [146] proposed a fuzzy emotion model that could analyze facial expressions in video sequences. Mandryk and Atkins [147] presented a knowledge-based fuzzy logic model using physiological data to recognize emotion. Arousal and valence values were generated by the physiological signals and were then used as inputs to a fuzzy logic model to detect emotional states including boredom, challenge, excitement, frustration, and fun.

Chatterjee et al. [148] used a data-driven method (one successful technique is neuro fuzzy [138]) to model face emotion by identifying the spots, edges and corners of a face and training the neuro fuzzy model. Ioannou et al. [149] have successfully proposed an emotion recognition system that analyzes and evaluates facial expressions incorporating psychological knowledge about emotion. A neuro fuzzy rule-based system has been created to classify facial expressions by analyzing facial animation parameter variations from the discrete emotional space and the continuous 2D emotion space. Katsis et al. [150] presented a methodology of assessing the emotional state of car-racing drivers using bio-signals. Neuro fuzzy and a support vector machine were used as the classification techniques. Lee et al. [151] and Giripunje et al. [152] used a data-driven based fuzzy inference system for emotion recognition from human speech.

2.5 Summary

This chapter introduces the basic approaches for categorizing emotions, and how to detect emotion through various methods for HCI. Based on the information conveyed by each modality, the emotion recognition can be achieved by four kinds of signals: audio, video, text, and physiological signals. We looked into these four kinds of signals and reviewed the recent works that have been done by applying each kind of single for interpreting emotions. The advantages and disadvantages of applying each modality are also discussed. For the multimodal approach in human emotion detection, fusion techniques and their pros and cons are also presented. Finally, the reviews of human behaviour recognition from body language of the head as well as emotion detection using soft computing techniques are given as well.

Chapter 3

Emotion Detection from Body Language of the Head

3.1 Introduction

Human body language of the head, especially eye gaze direction and head movement direction as well as the frequency of head movement have been ignored in the past research in the process of recognizing emotion. Most existing approaches in human computer interactions for emotion detection have focused on the recognition of the single facial expression modality [130], [131] [23]. However, detecting facial expression, especially the six basic facial expressions (happiness, sadness, surprise, angry, disgust, and fear), is only the tip of the iceberg for emotion recognition. Emotion expression is performed through simultaneous combination of various nonverbal cues. Attempting to infer emotion from facial expression alone does not always lead to an accurate conclusion. In particular, psychological research shows that the direction of eye gaze and head movement has a strong effect on facial expression [127], [128], [129]. When a facial

expression is accompanied by other nonverbal cues, entirely different emotions may be inferred. For example, the usual interpretation of smiling facial only is not necessarily correct when the person is also shaking his or her head. It is more likely that he or she is not happy. Therefore, considering other modalities is crucial for analyzing emotions. Moreover, there are far more secondary and tertiary emotions than the basic facial expressions. Recognizing these subtle secondary emotions is critical for HCII applications. Additionally, researchers have shown that cognitive mental states (e.g., agreement, disagreement, concentrating, thinking, and interest) occur more frequently in daily interactions than the six basic facial expressions [13]. Despite the importance of body language of the head in interpreting emotions, none of the previous works have considered combining these modalities to infer emotions.

The goal of this work is to explore new ways of HCII by integrating information from the body language of the head to infer emotions and cognitions. A two stage approach is proposed. In the first stage, the explicit information from body language of the head is detected. In the second, stage, the multimodal information is fused to infer implicit human complex emotional states. In particular, we concentrated on the combination of facial expression, eye gaze, and head movement. This is a difficult task since it requires the incorporating of knowledge from various disciplines, especially psychology.

Soft computing techniques [136], [137], [138] are applied since emotions have fuzzy boundaries [54]. The multimodal fusion is done by hybridizing the decision level fusion and the soft computing techniques for classification. This could avoid the disadvantages of the decision level fusion technique, while retaining its advantages of adaptation and flexibility. Fuzzification strategies are proposed which can successfully quantify the extracted parameters of each modality into a fuzzified value between 0 and 1. These fuzzified values are the inputs for the fuzzy inference systems (FIS) [136], [137], [138] which map the fuzzy values into emotional states.

3.2 Fuzzy System for Emotion Detection

Effective design of emotion recognition systems is a very challenging task since it relies on multidisciplinary collaboration among areas such as computer science, electrical engineering, and psychology for the sharing of knowledge. Moreover, emotion detection

is difficult since the boundaries among different emotions are not crisp [54]. For example, sometimes different emotions have the same facial expression. In this thesis, fuzzy inference systems [136], [137] are applied for emotion and cognition detection.

The general model of emotion detection process is shown in Figure 3.1. First of all, we extract information from the body language of the head. This extracted information is used as input for the fuzzification interface, which defines a mapping from a crisp value to a fuzzy number. These fuzzy numbers are then given to the knowledge-based inference rules, which give conditions to derive reasonable actions. The defuzzification interface then maps the fuzzy value into a crisp emotional or cognitive state.

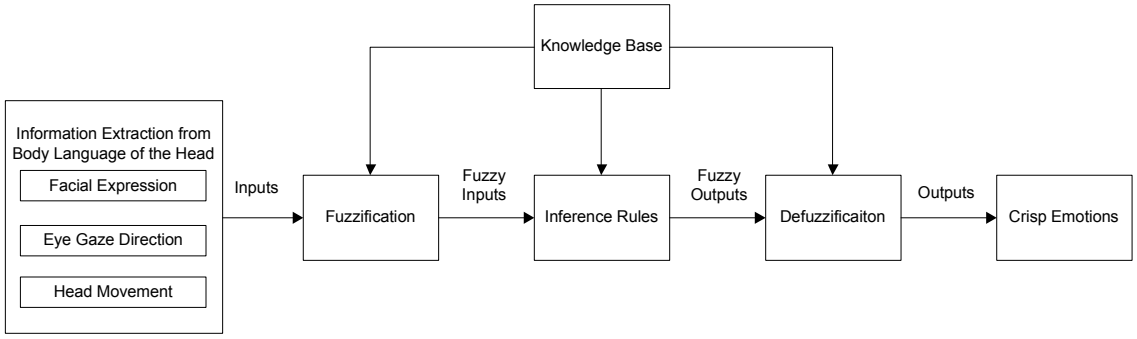


Figure 3.1 General model of emotion detection

3.2.1 Input Variables

In this sub-section, the input variables used in the fuzzy system are discussed. Figure 3.2 shows the taxonomy structure of the body language of the head. For the fuzzy inference systems, we defined five input linguistic variables: Happiness, Angry, Sadness, Head Movement, and Eye Gaze (see Figure 3.3 to Figure 3.7). The input variable "Happiness" was mapped into membership functions as "Maybe Happy", "Little Happy", "Happy", and "Very Happy" according to the degree of happiness. The input variable "Angry" was mapped into membership functions as "Maybe Angry", "Angry", and "Very Angry" based on the degree of angriness. The input variable "Sadness" was mapped into membership functions as "Maybe Sad", "Sad", and "Very Sad" based on the degree of sadness. The input variable "Head Movement" was mapped into membership function as "High Frequency Shaking", "Low Frequency Shaking", "Stationary", "Low Frequency Nodding", and "High Frequency Nodding" based on the direction and frequency of head movement. The input variable "Eye Gaze" was mapped into membership function as "Avert Gaze" and "Direct Gaze" based on the scale of eye gaze direction. The mapping

strategy from input variables to membership function boundary parameters will be explained in the experiment and results section.

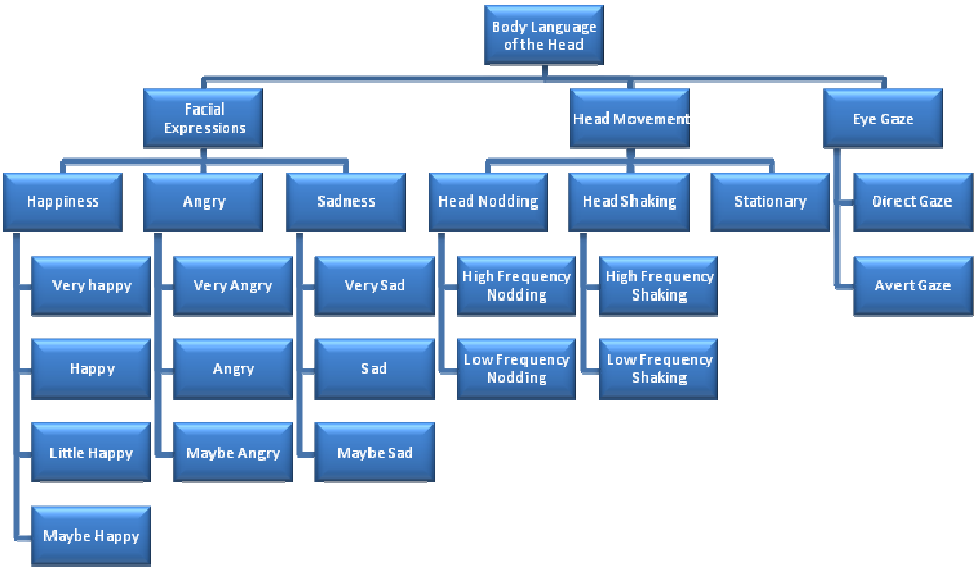


Figure 3.2 Taxonomy structure of body language of the head

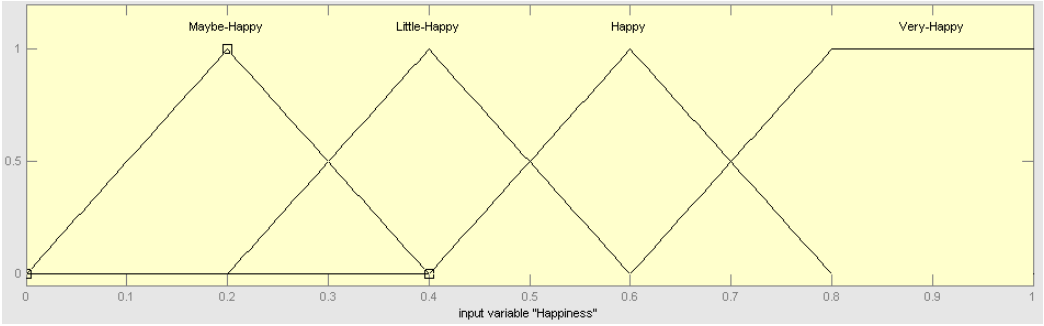


Figure 3.3 Mamdani-type input happiness variables

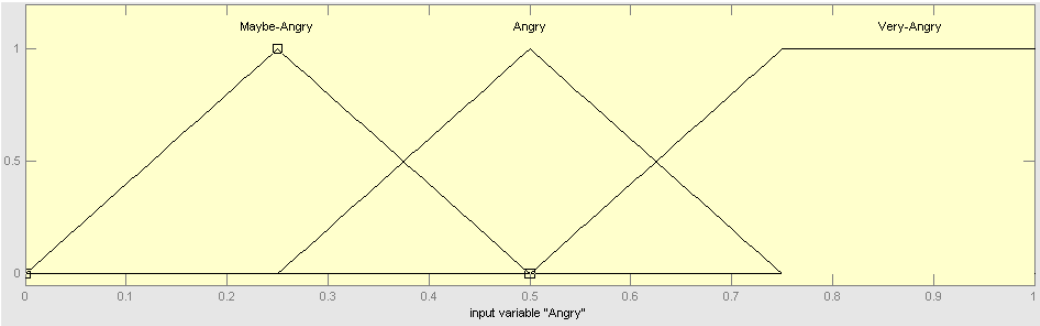


Figure 3.4 Mamdani-type input angry variables

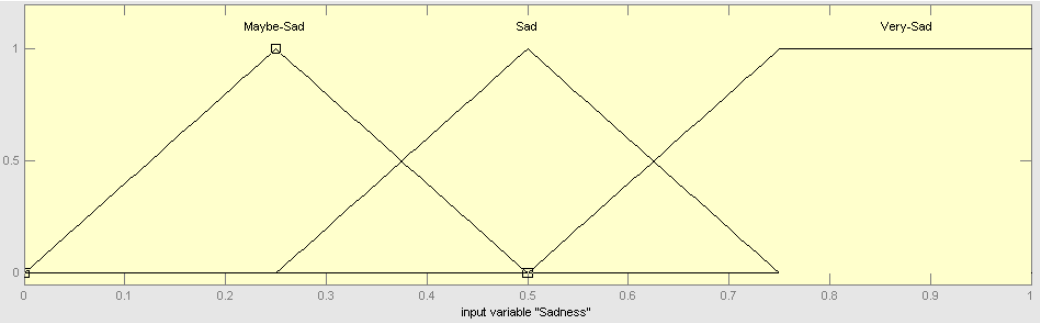


Figure 3.5 Mamdani-type input sadness variables

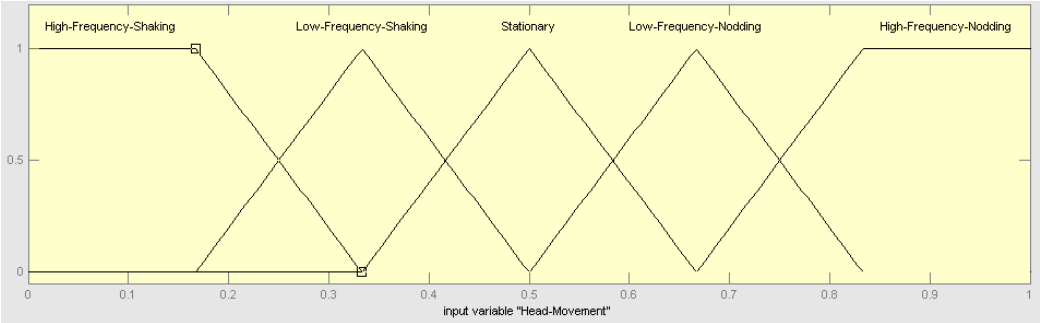


Figure 3.6 Mamdani-type input head movement variables

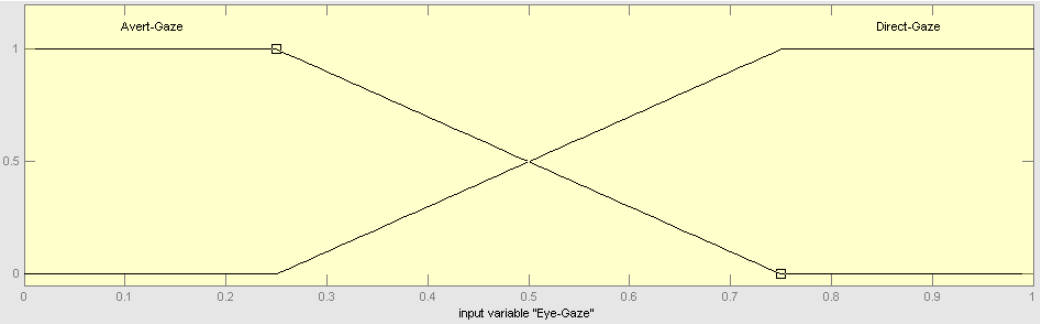


Figure 3.7 Mamdani-type input eye gaze variables

These linguistic variables need to be further quantified into fuzzy input values. The variables can be quantified into input values ranging in degree between 0 to 1 using the following proposed method respectively:

1) We quantified the linguistic variable “Head Movement” into values using the following formulas:

$$\begin{aligned} &\text{Head-movement} = 0.5 + (\text{sgn}) \times f_{normal} \\ &\text{sgn} = \begin{cases} 1\ , nodding \\ -1, shaking \end{cases} \\ &f_{normal} = \frac{f_{hm}}{2 \times MaxFrequency}, \quad hm = nod \text{ or } shake \end{aligned} \tag{3.1}$$

where Head-movement is the quantified input value, f_{hm} is the frequency of a certain head nod or head shake movement, and *MaxFrequency* is the maximal frequency of head nods and shakes. After repeated testing, we found that the maximum frequency human beings can achieve for both head nods and shakes is 3 cycle/sec. Therefore, 3 would be a reasonable value for *Maxfrequency*. The detail steps for acquiring the head movement direction and the movement frequency will be discussed in Chapter 4.

2) We computed the value of S for the scale of eye gaze direction analysis, which is defined in formula 5.2 (detail steps will be discussed in Chapter 5). The input value of the “Eye Gaze” variable can be defined using the following formula:

$$EyeGaze = 1 - \bar{S}_{normal} \tag{3.2}$$

where $\bar{S}_{normal}$ is the normalization value *S*. The detailed procedures will be explained in the experiment section.

3) For the facial expression analysis, for each three category of facial expression, we trained a classifier that divided expressions by intensity. The detail steps of how to train the classifier will be explained the Chapter 6. The classification results will then manually marked with values.

3.2.2 Establishing Fuzzy Rules

In order to obtain valid emotions to generate fuzzy rules, the ideal conditions for the experiment should be that:

1) The subject feels the emotion internally.
2) The subject should be in a real-world environment instead of a lab environment, and emotions should occur spontaneously.
3) The subject should not be aware that he or she is being recorded.
4) The subject should not know that he or she is part of an experiment.

The ideal experiment cannot be conducted due to privacy and ethical concerns. Therefore, we used an alternative method for obtaining ground truth and fuzzy rules from the knowledge and opinions of an expert in psychology, a pilot group, and annotators (as shown in Figure 3.8). The generation procedure of the culturally determined display rules are as follow:

Step 1: A psychology expert gave advice on high-level instructions (for example, emotions can be inferred by a combination of facial expressions, eye gaze direction, and head movements). These instructions were used for the pilot group, which consisted of two Ph.D. students (one majoring in Computer Science and the other in Electrical Engineering) and one undergraduate student (majoring in Psychology). The pilot group made decisions about candidate rules. After discussion, 15 candidates for fuzzy rules (each corresponding to an emotion or cognition) were defined.

Step 2: Fifteen video clips representing these 15 emotions and cognitions were recorded by one member of the pilot group. Self-report questionnaire-type scales have been extensively used to assess emotions. We established a questionnaire to evaluate the proposed rules. In the questionnaire, the following question was asked: “After watching this video clip, what emotion or cognition have you experienced?” for each video clip.

Step 3: Ten annotators, who were not members of the pilot group, were asked to annotate these video clips by answering the questionnaire. When answering the questionnaire, some annotators used alternative words to depict emotion. For example, for “resentment,” some annotators used “hate.” For the “decline” response, some annotators used “reject politely” instead. In such cases, we considered that both terms represented the same emotion. In order to minimize the influence of out-group culture differences (in-group members have an advantage in recognizing emotions [153]), we asked only Chinese students to be annotators in this experiment.

Step 4: The pilot group analyzed the answers in the questionnaire. A fuzzy rule would be generated if the following condition is satisfied: at least 8 out of 10 annotators have experienced the same emotion from watching the clip. After analysis, 11 out of 15 candidate fuzzy rules were retained. The other four were rejected.

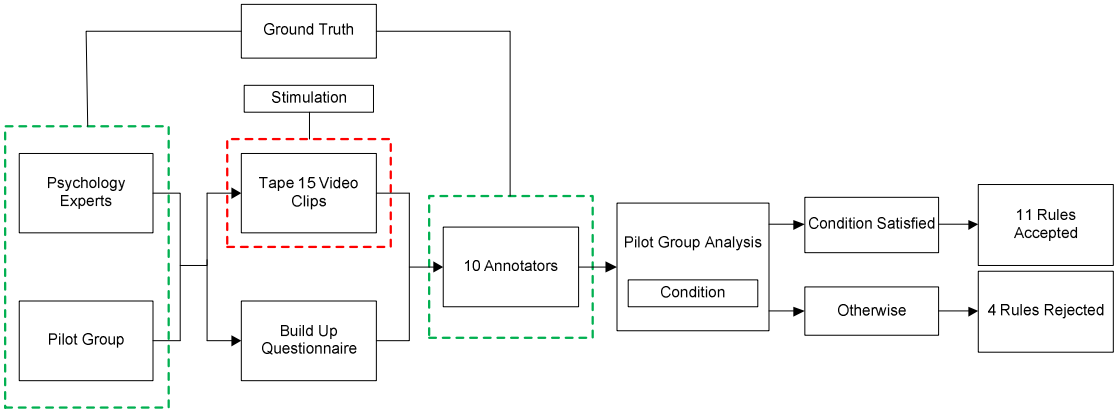


Figure 3.8 Procedure of fuzzy rules creation

It is worth mentioning that although emotions are universal, the ways humans express emotions vary between cultures [154], [155], [156]. Interpreting human emotions or cognitions in the context of normal body language sometimes leads to misunderstandings and misinterpretations. People from different cultures can interpret body language in different ways. For example, nodding may mean different messages in various parts of the world. In China, it means *I agree*. But in some cultures, like parts of India, Bulgaria, and Turkey, head nods means *no*. In a conversation among Japanese, it simply means *I am listening*. In most cases, head nodding is associated with positive emotions, whereas head shaking is associated with negative emotions. However, for some individuals, high-frequency head nodding is likely to occur with resentment, which is a negative emotion. Different cultures also have different rules for eye contact. Some Asian cultures may perceive direct eye contact to be inappropriate. In some countries, lower one's eyes show a signal of respect, while similarly eye contact is avoided in Nigeria [157]. On the other hand, in western cultures this could be misinterpreted as lacking confidence.

To the best of our knowledge, only the six basic facial expressions (happiness, anger, fear, surprise, sadness, and disgust) have been proven by psychologists to be universal. This is the reason why we asked students with the same cultural backgrounds to participate in this experiment. Although the fuzzy rules were drawn by in-group people,

the idea of integrating the different modalities of the body language of the head is generic enough to be used by any particular target user group from any culture. For example, if we want to recognize emotions with people from a different culture, this could be accomplished by simply modifying the fuzzy rules of the fuzzy inference system. The input information of the three modalities remains the same.

Regarding the ground truth, some facial emotion researchers use movie clips to stimulate emotions. However in our case, emotions were complex and many (such as guilt) could not be easily evoked by movie clips or other methods. Therefore, video clips were taped and shown to the annotators in an effort to provoke emotion and cognition. These video clips play the same role as movie clips to provide the stimulation. The 11 generated rules based on the above procedures are listed below:

Rule 1: IF (*Happiness* is *Very-Happy*) AND (*Head-Movement* is *High-Frequency-Nodding*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-A* is *Agree*)

Rule 2: IF (*Happiness* is *Happy*) AND (*Head-Movement* is *Low-Frequency-Nodding*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-A* is *Admire*)

Rule 3: IF (*Happiness* is *Little-Happy*) AND (*Head-Movement* is *Low-Frequency-Shaking*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-A* is *Decline*)

Rule 4: IF (*Happiness* is *Maybe-Happy*) AND (*Head-Movement* is *Low-Frequency-Nodding*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-A* is *Thinking*)

Rule 5: IF (*Happiness* is *Maybe-Happy*) AND (*Head-Movement* is *Stationary*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-A* is *Thinking*)

Rule 6: IF (*Angry* is *Very-Angry*) AND (*Head-Movement* is *Low-Frequency-Shaking*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-B* is *Resentment*)

Rule 7: IF (*Angry* is *Maybe-Angry*) AND (*Head-Movement* is *High-Frequency-Shaking*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-B* is *Disagree*)

Rule 8: IF (*Sadness* is *Very-Sad*) AND (*Head-Movement* is *Low-Frequency-Shaking*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-C* is *Distressed*)

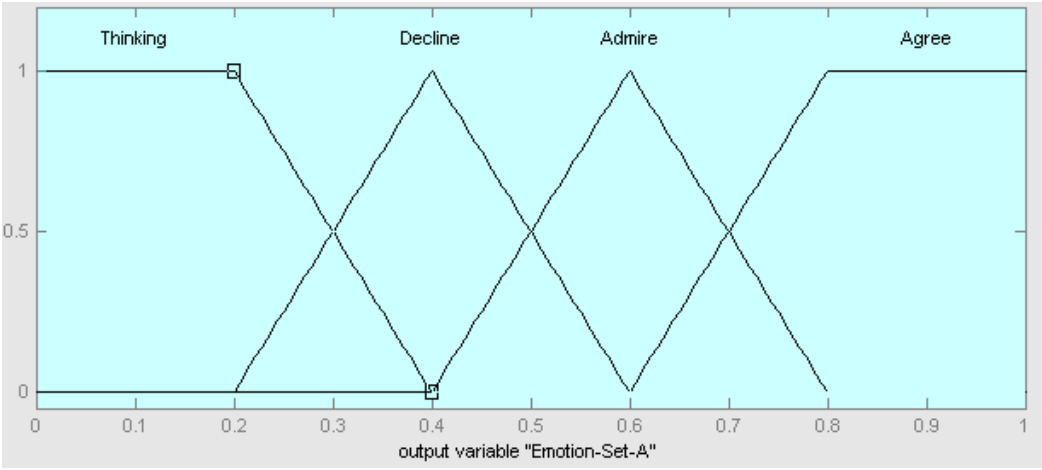
Rule 9: IF (*Sadness* is *Sad*) AND (*Head-Movement* is *High-Frequency-Nodding*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-C* is *Guilty*)

Rule 10: IF (*Sadness* is *Maybe-Sad*) AND (*Head-Movement* is *Low-Frequency-Shaking*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-C* is *Disappointed*)

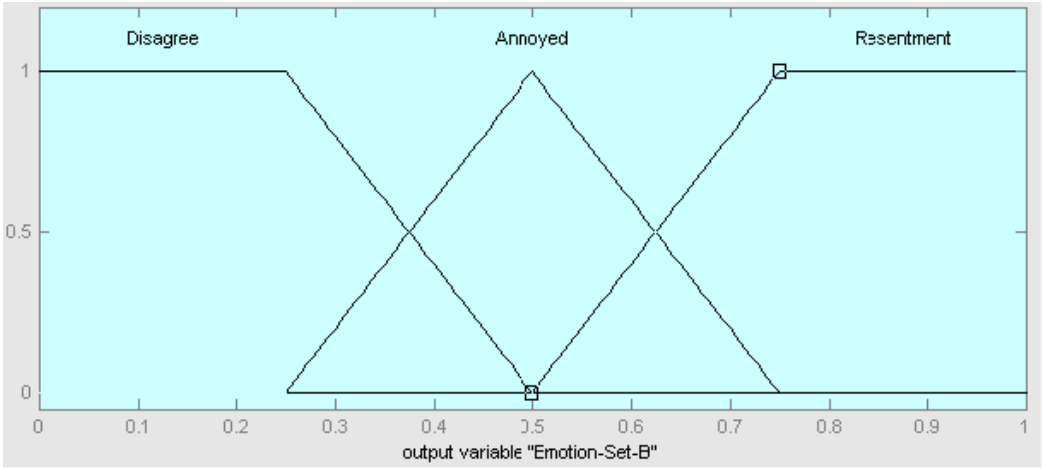
Rule 11: IF (*Angry* is *Angry*) AND (*Head-Movement* is *High-Frequency-Shaking*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-B* is Annoyed)

3.2.3 Output Variables

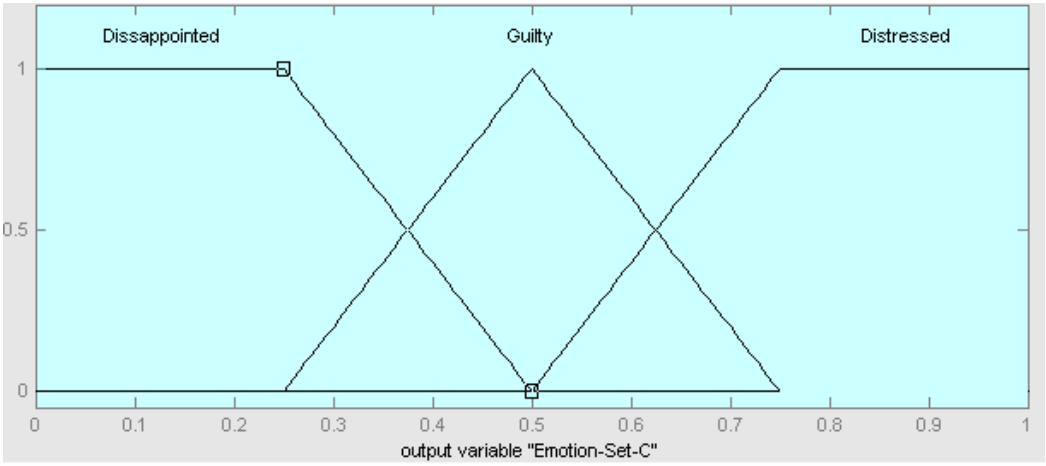
The output variables from our emotion detection system are Emotion set-A, Emotion set-B, and Emotion set-C. From the rules described above, we came up with 10 emotions and cognitions. These emotions and cognitions can be divided into three categories, which are derived from the three facial expressions: happiness, angriness, and sadness respectively. Emotion set-A includes “Thinking”, “Decline”, “Admire”, and “Agree”, which are co-related with the happy expression. Emotion set-B includes “Disagree”, “Annoyed”, and “Resentment”, which are co-related with the angry expression. Emotion set-C includes “Disappointed”, “Guilty”, and “Distressed”, which are co-related with the sad expression. Therefore, we set up three output variables as shown in Figure 3.9.



a) Output variable Emotion set-A



b) Output variable Emotion set-B



c) Output variable Emotion set-C

Figure 3.9 Output variables and their respective membership functions

3.3 Experiment and Results

3.3.1 Database

The database of emotion samples used in the experiment is collected from a group of five Chinese university students (aged from 20 to 35), including three females and two males, volunteered to participate in the experiment. The experiment was carried out in a laboratory environment with uniform illumination. A commercial-grade 640*480 pixel Microsoft LifeCam (see Figure 3.10) was used to record all the videos at 15 frames/s for approximately 5 seconds each. The participants were asked to perform the emotions and

cognitions twice according to the 11 established rules. Thus, a total of 110 video clips were recorded. Then, each video clip was divided into image sequences with time intervals of 100ms. We extracted 30 consecutive frames from each image sequence, manually choosing the first frame that showed a frontal face. These 110 image sequences were used as a database for input value extractions for fuzzy inference systems.

Figure 3.10 Microsoft LifeCam Web Camera

3.3.2 The performance experiments

In the experiment, both conventional fuzzy logic (a knowledge-based technique that manually annotates the mapping strategy from input values to membership function boundaries) and neural fuzzy techniques (a data-driven-based technique that obtains membership function boundaries from training data) were utilized for emotional and cognitional detection. In order to evaluate both Mamdani-type FIS and Sugeno-type FIS, we tested the recognition rate of the 110 image sequences for each system.

3.3.2.1 Fuzzy logic system performance

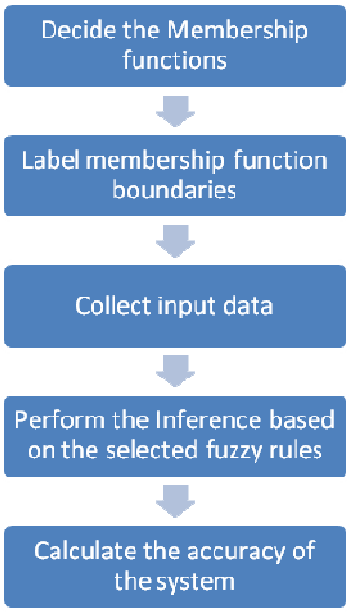


Figure 3.11 The steps of fuzzy logic based approach

A conventional Mamdani-type [160] fuzzy inference system is an intuitive technique that has the main advantage of being capable of formalizing inputs and fuzzy rules from expert knowledge, especially in cases in which there is no sufficient information available, where using expert knowledge is the only way to collect information [136]. The general steps of fuzzy logic based approach are shown in Figure 3.11. Two typical triangular and trapezoidal membership functions were applied. We manually labelled membership function boundary parameters uniformly based on the assumption that the possibility of each case is uniformly distributed.

The steps for obtaining fuzzy input values for eye gaze were as follows:

1. We extracted the frontal face images from the 30 images using Formula 5.3.
2. For each frontal face image, the S value was calculated using Formula 5.2, and then the average S value was calculated for all frontal images.
3. By applying Formula 3.2, the fuzzy input could be obtained for the image sequence.
4. For all image sequences, repeat the above steps.

The steps for obtaining fuzzy input values for head movement were as follows:

1. The nostril points for each image were extracted, tracked, and recorded for an image sequence based on the methods in Chapter 4.
2. Afterwards, head movement direction and frequency were analyzed based on the

algorithms in Chapter 4.

3. Using Formula 3.1, we obtained the fuzzy input for the image sequence.
4. For all image sequences, repeat the above steps.

The steps for obtaining fuzzy input values for facial expression were as follows:

1. We extracted the frontal face images from all of the 110 images using Formula 5.3.
2. We manually categorized them into three emotion sets: happiness, sadness, and angry. For each emotion set, using the methods provided in Chapter 6, train each category of emotion set into discrete degree of emotion, and labelled the output into values between 0 to 1.
3. For each image sequence, we first classify to a specific emotion set, and then classify the degree of emotion.
4. For all image sequences, repeat the above steps.

Figure 3.12 shows an image sequence example expressing the “Admire” emotion. Table 3.1 lists the example of the input and output values according to this image sequence. The quantified input values can be obtained from the formulas defined in section 3.2.1. For example, after calculating the frequency value using FFT (0.8 cycle/sec) and recognizing the head movement direction (nodding), formula (3.1) can be applied to determine the input value of head movement, which is 0.633. Following the same strategy, we obtained the input value for eye gaze and facial expression. By applying the knowledge-based Mamdani-type fuzzy inference system, which manually labels membership function boundary parameters uniformly, the output value is 0.6, which is the “Admire” emotion. Only rule 2 has been activated (see Figure 3.13).

Figure 3.12 An example of an image sequence for Admire emotion

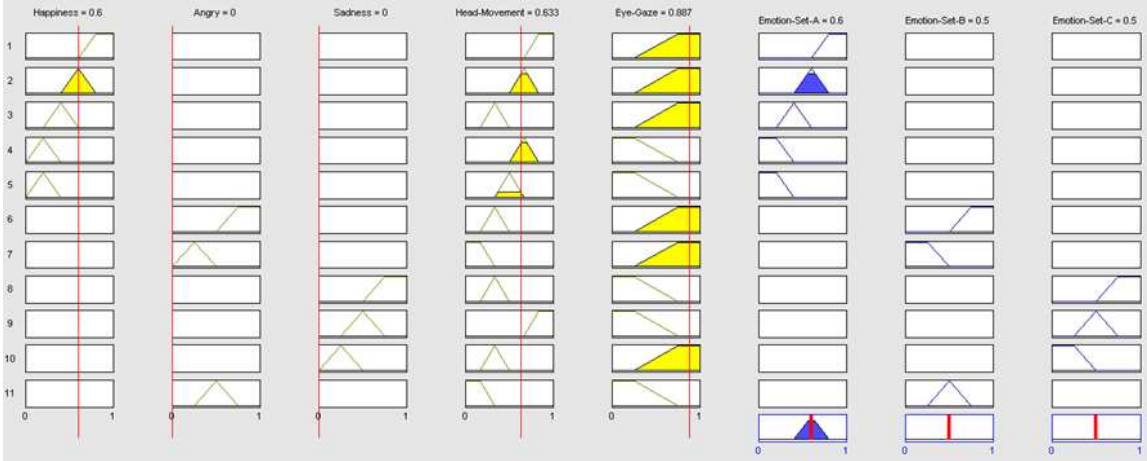


Figure 3.13 View of rule for Admire emotion

Table 3.1 An example of the input and output values

The results of each stage	Head movement		Eye gaze	Facial expression	Emotion	
	Status Nodding	Frequency 0.8 cycle/s	Direct	Happy	Output	Emotion
Value of input fuzzy variable	0.633		0.887	0.6	A = 0.6 B = N/A C = N/A	Admire

3.3.2.2 Neuro fuzzy system performance

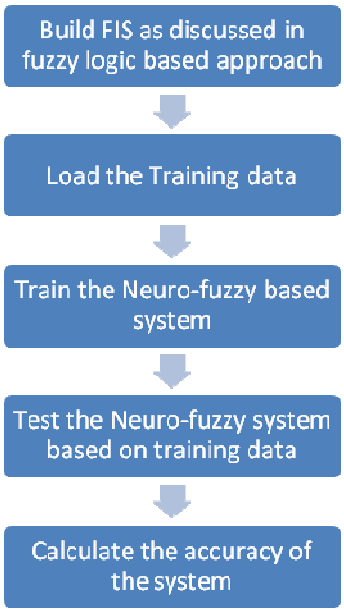


Figure 3.14 The general steps of Neruo fuzzy based approach

A hazard of conventional Mamdani-type fuzzy systems is that it may be based on qualitative and imprecise knowledge from an operator's or expert's perspective. It requires a thorough understanding of the fuzzy variables, their mapping strategy, and their membership functions, as well as the selection of fuzzy rules. In contrast, neural fuzzy systems can enhance fuzzy reasoning by acquiring knowledge from training input data. This has the advantage that it lets the data "speak" for itself. The general steps of Neruo fuzzy based approach is shown in Figure 3.14. To take advantage of that, we also applied a Sugeno-type [159] neuro fuzzy system that can generate membership function boundaries automatically from training data.

Figure 3.15 shows an example of view of rule for the neuro fuzzy system. From the input variables of each modalities (*Angry* is *Maybe-Angry*, *Head-Movement* is *High-Frequency-Shaking*, and *Eye-Gaze* is *Direct-Gaze*), by applying the Sugeno-type neuro fuzzy system, the crisp output belongs to Emotion-set B which is *Disagree*.

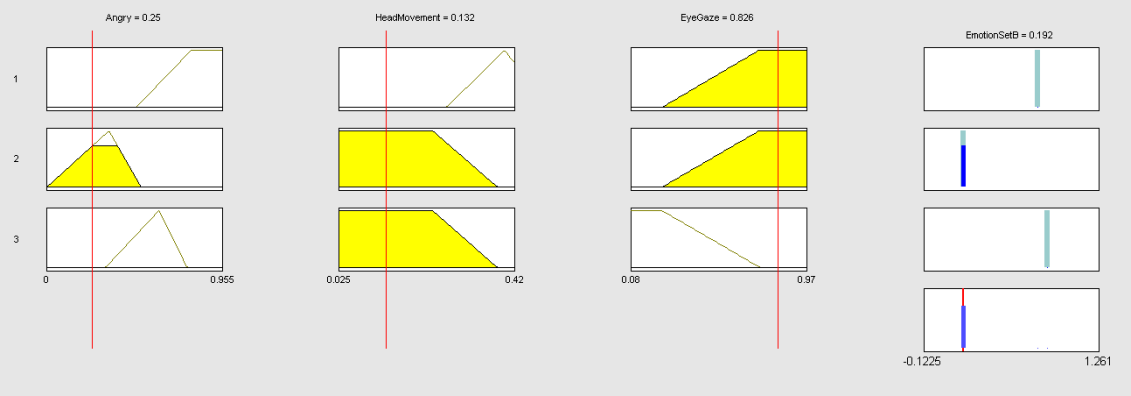


Figure 3.15 View of rule for Neuro Fuzzy

Leave-one-out validations

The performance was tested by using the leave-one-out cross validation procedure, in which one participant's data was withheld from the FIS training and retained for testing, and the test was repeated for each participant. Hybrid [138] provided the optimal method for defining the parameters of the input and output membership functions. After training, we calculated the average testing error between the trained FIS output and the testing data for each participant.

Figure 3.16 to Figure 3.30 show the comparison between the Neuro FIS outputs and the testing data outputs. Since there are 5 participants and 3 sets of emotions, therefore,

there are 15 figures in total. The x-axis is the number of times the experiments were performed for a particular emotion set. Take Figure 3.16 as an example, 6 times of experiments (3 rules for emotion-set-B, each rule tested twice) have been tested for participant 5 for emotion-set-B. The y-axis stands for the corresponding fuzzy output after each experiment. The fuzzy inference systems then maps the fuzzy output into a crisp emotion as discussed in Section 3.2.3 and Section 3.3.2.1. The red asteroids represent the FIS outputs and the blue dots represent testing data outputs. Table 3.2 summarized the average testing errors of the neural fuzzy classifier.

3.3.2.3 Results and discussions

In the experiments, all of the fuzzy outputs are mapped into crisp emotions. If the resulting emotion is different from the true emotion, the recognition result is incorrect. Table 3.3 shows the recognition results of each method under different emotional sets. The reasons for the errors are as follows: (1) The input value for the fuzzy inference system is not correct. For example, if the gaze direction has been incorrectly recognized, the corresponding fuzzy rule will not be activated correctly, resulting in the failure of final emotion recognition. (2) In neuro fuzzy system, if the testing sample is very different from the training data sets, this will also result in wrong emotion output.

In use, the choice between the two types of fuzzy inference systems will depend on practical considerations. The membership function boundaries for knowledge-based Mamdani-type FIS are manually labelled, assuming that the cases are uniformly distributed. However, there is a chance that these boundaries are not perfectly uniformly distributed for a particular target user group. Therefore, we further applied the Sugeno-type neural FIS which could automatically generate membership functions boundaries based on the training data sets. When the training data is insufficient, the disadvantage of a Sugeno-type fuzzy system is that the training data set might not fully represent the target group and thus leads to incorrect recognition. On the other hand, the Mamdani-type system is easy and intuitive, but less adaptive to changes. The neural fuzzy system can be improved by training with more input and output data, while the Mamdani-type system cannot. Thus, if more input and output data are collected from experiments in the future, the recognition rate of the neural fuzzy system can be improved. The advantages of Sugeno-type method and Mamdani-type method are summarized as follow:

- Advantages of Sugeno-type method
  1. It is computationally efficient
  2. It works well with linear techniques
  3. It works well with optimization and adaptive techniques
- Advantages of Sugeno-type method
  1. It is intuitive
  2. It has widespread acceptance
  3. It is well suited to human input

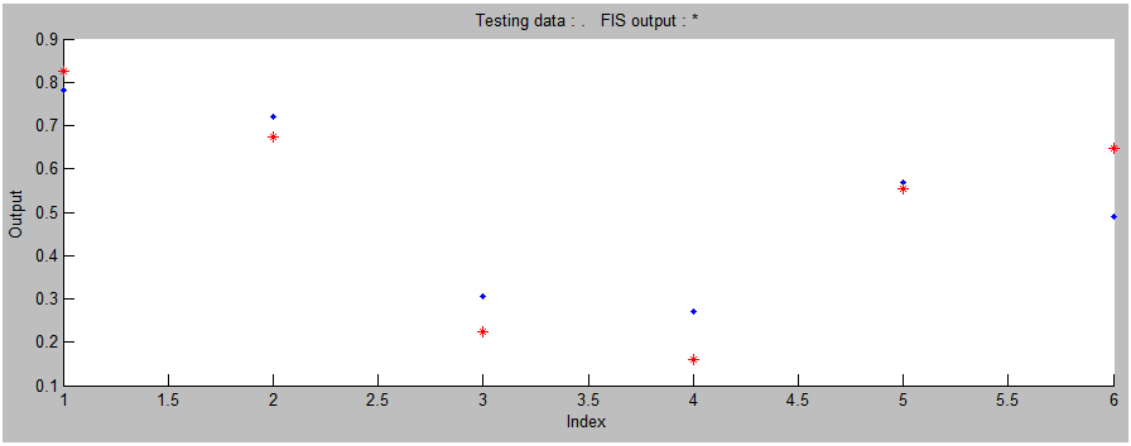


Figure 3.16 The output comparision between FIS output and Testing data of participant 5 for Emotion-Set-B

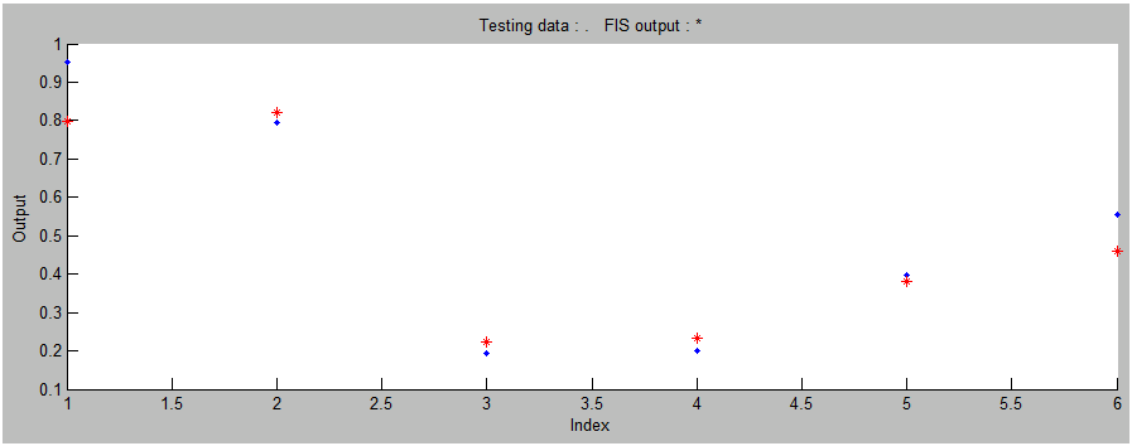


Figure 3.17 The output comparision between FIS outputs and Testing data of participant 4 for Emotion-Set-B

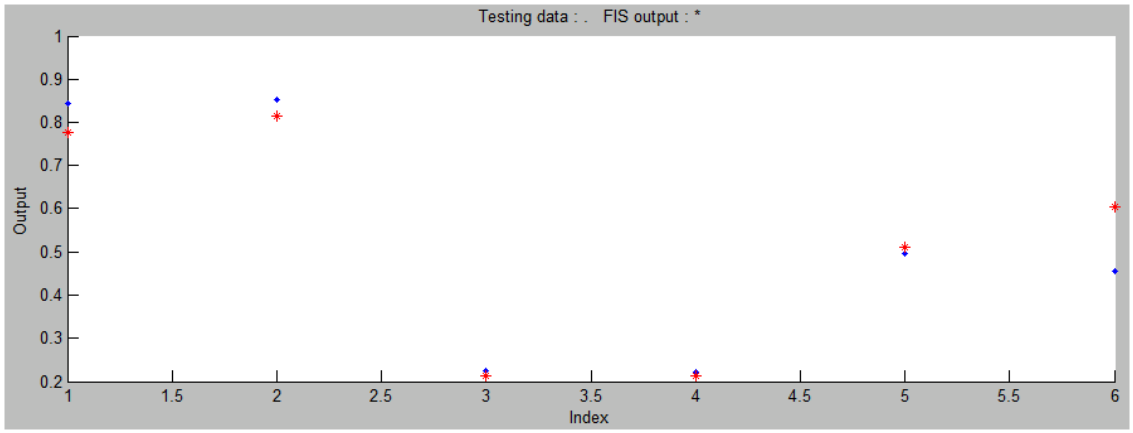


Figure 3.18 The output comparision between FIS outputs and Testing data of participant 3 for Emotion-Set

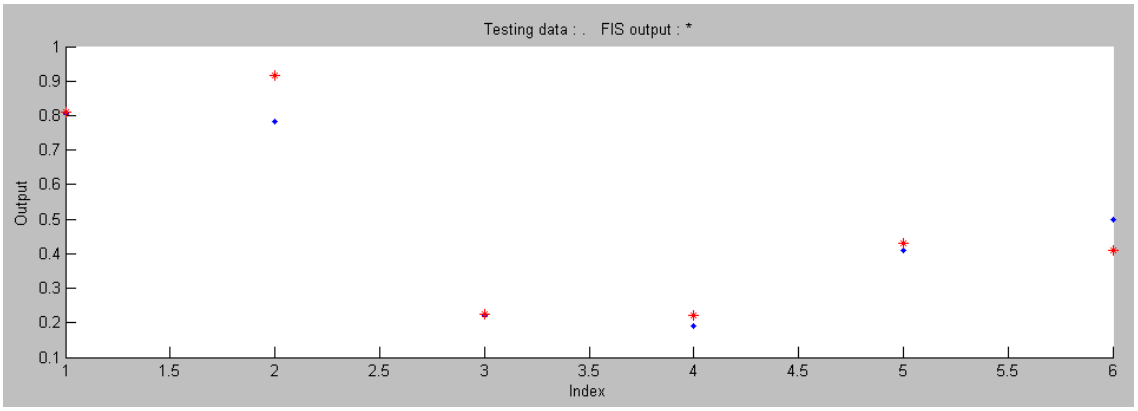


Figure 3.19 The output comparision between FIS outputs and Testing data of participant 2 for Emotion-Set-B

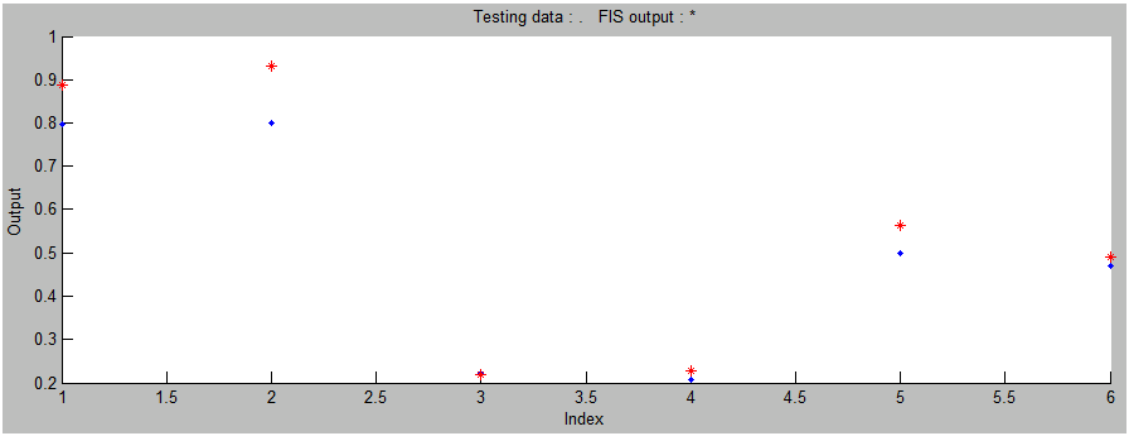


Figure 3.20 The output comparision between FIS outputs and Testing data of participant 1 for Emotion-Set-B

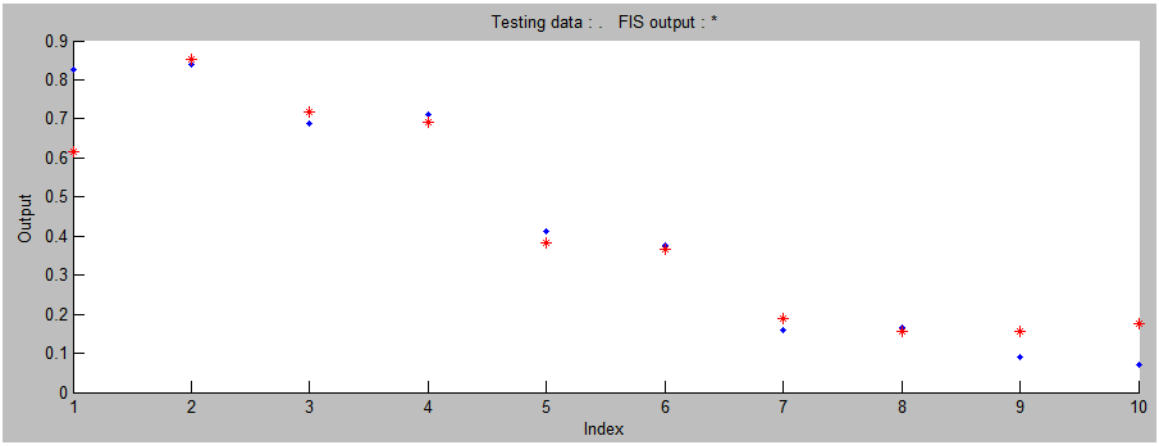


Figure 3.21The output comparision between FIS outputs and Testing data of participant 5 for Emotion-Set-A

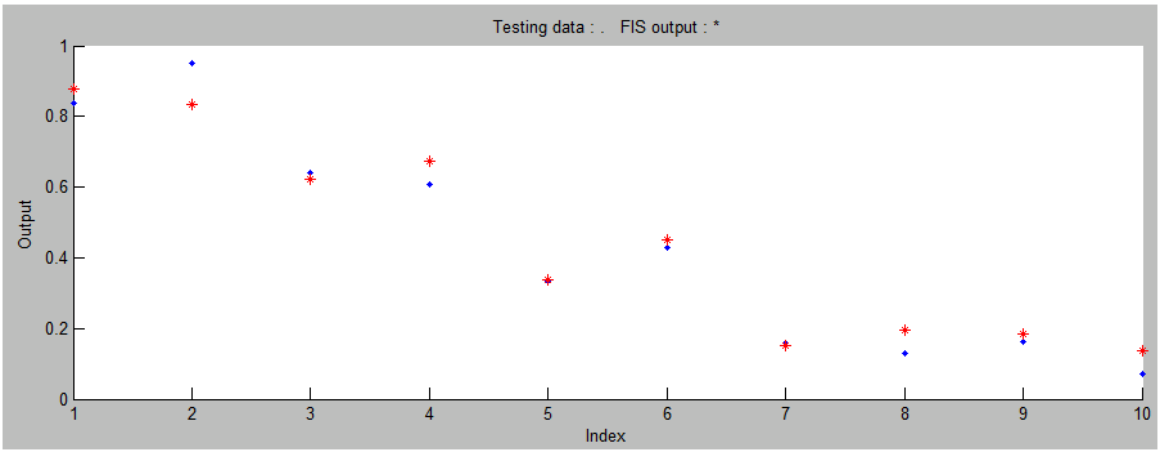


Figure 3.22 The output comparision between FIS outputs and Testing data of participant 4 for Emotion-Set-A

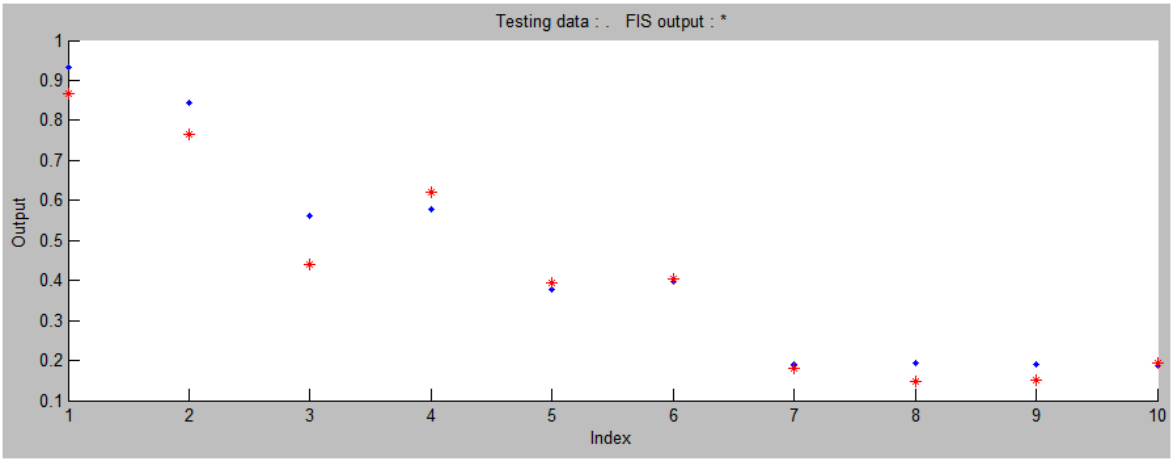


Figure 3.23 The output comparision between FIS outputs and Testing data of participant 3 for Emotion-Set-A

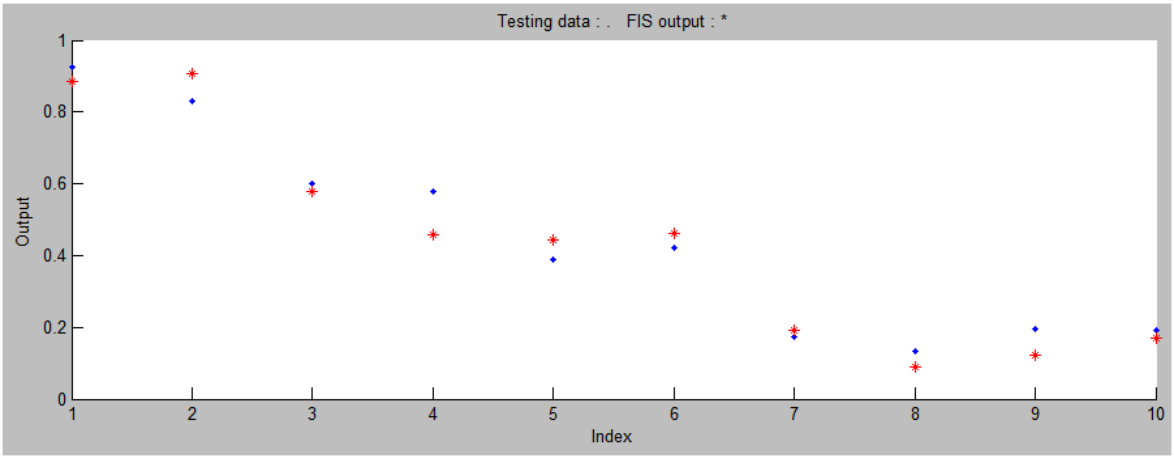


Figure 3.24 The output comparision between FIS outputs and Testing data of participant 2 for Emotion-Set-A

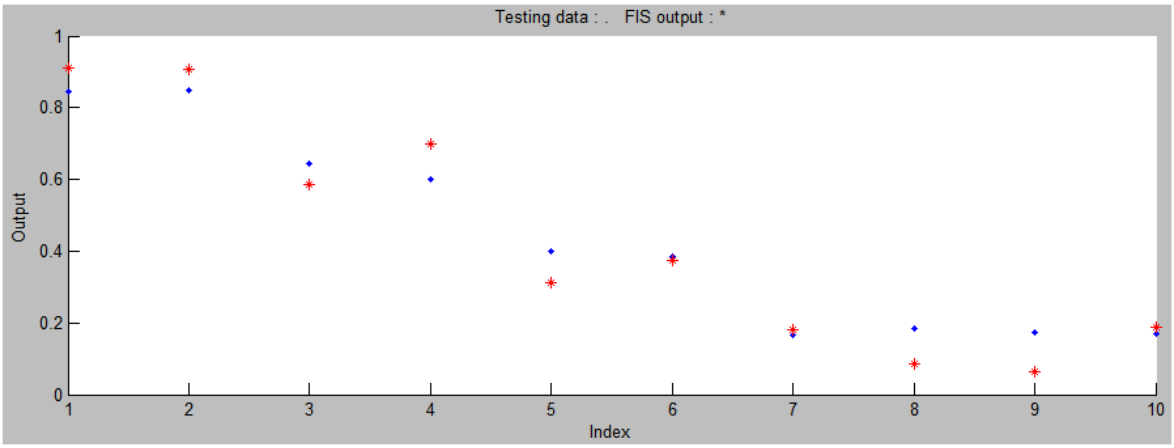


Figure 3.25 The output comparision between FIS outputs and Testing data of participant 1 for Emotion-Set-A

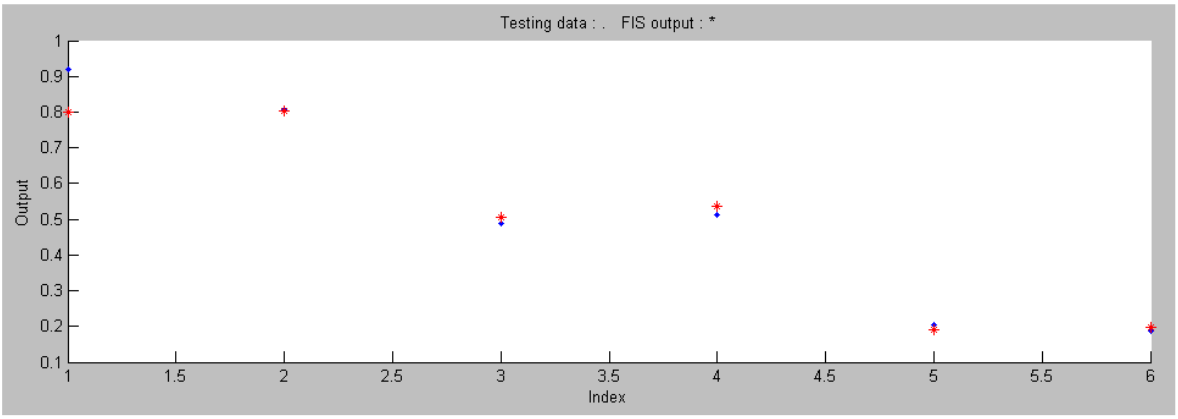


Figure 3.26 The output comparision between FIS outputs and Testing data of participant 5 for Emotion-Set-C

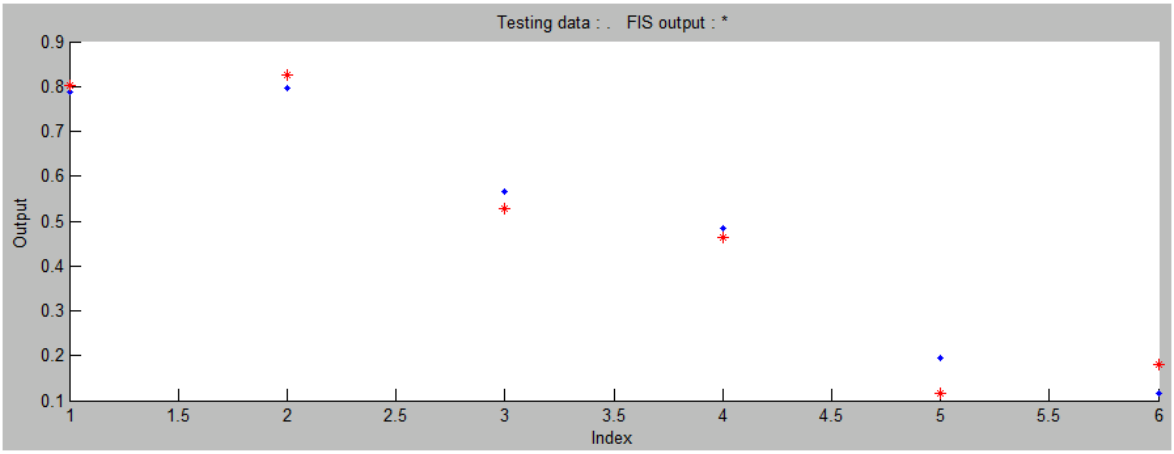


Figure 3.27 The output comparision between FIS outputs and Testing data of participant 4 for Emotion-Set-C

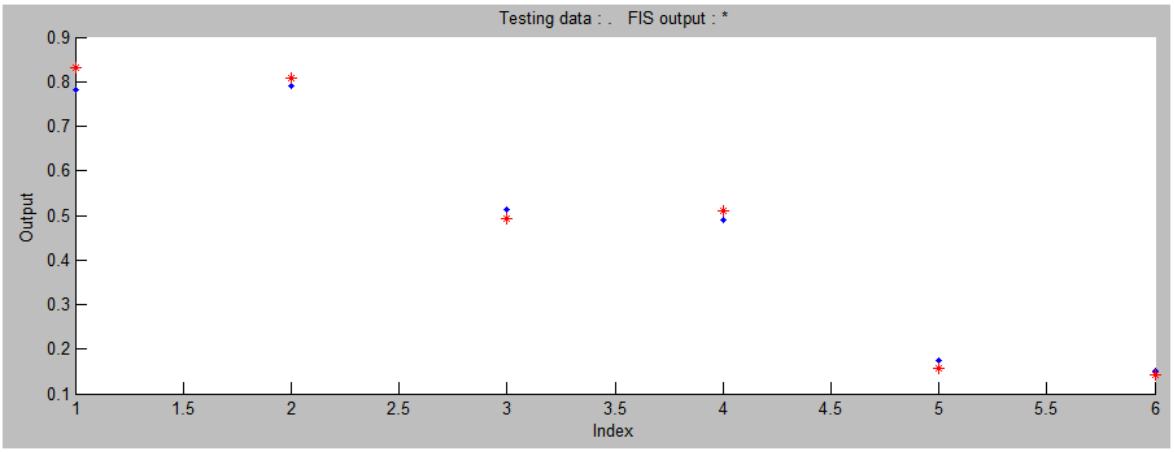


Figure 3.28 The output comparision between FIS outputs and Testing data of participant 3 for Emotion-Set-C

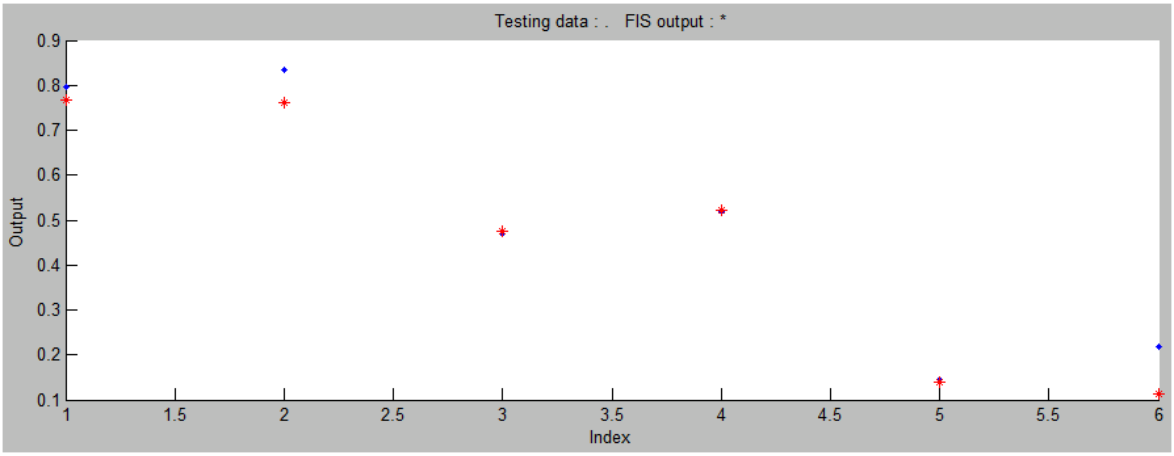


Figure 3.29 The output comparision between FIS outputs and Testing data of participant 2 for Emotion-Set-C

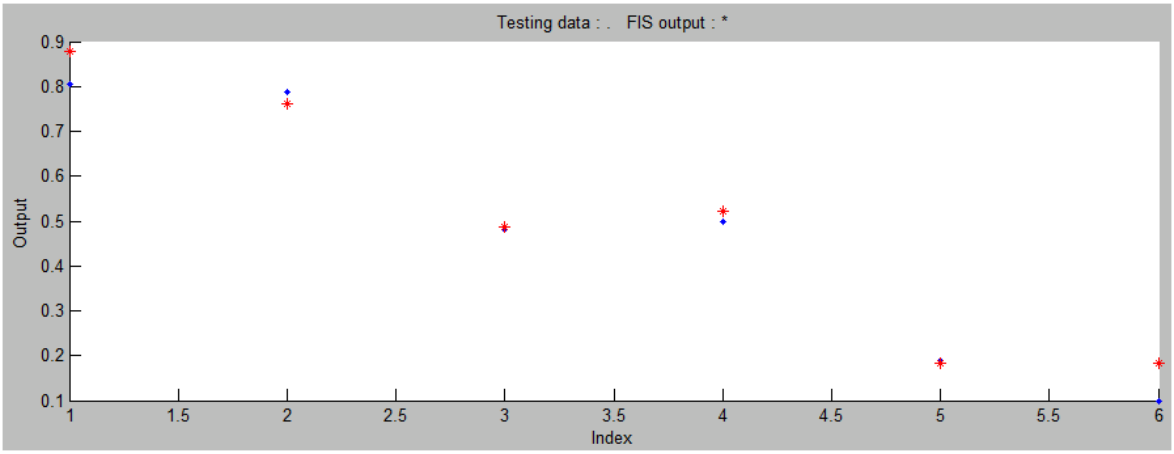


Figure 3.30 The output comparision between FIS outputs and Testing data of participant 1 for Emotion-Set-C

Table 3.2 Average Testing Error

Participants	Average Testing Error		
	Emotion Set A	Emotion Set B	Emotion Set C
1	0.071485	0.080826	0.04722
2	0.059401	0.22968	0.053492
3	0.05659	0.06852	0.025724
4	0.054581	0.076242	0.04675
5	0.078943	0.089649	0.052105
Average	0.0642	0.1089834	0.0450582

Table 3.3 Recognition Rate Comparison

	Recognition Rate		
	Emotion Set A	Emotion Set B	Emotion Set C
Sugeno-type	96%	93.32%	96.66%
Mamdani-type	96%	96.66%	96.66%

3.4 Summary

This chapter introduced new approaches of recognizing human emotions from the body language of the head. A two-stage approach is proposed. The first stage analyzes the explicit information from the individual modalities of facial expression, head movement, and eye gaze. In the second stage, the multimodal information is fused to infer implicit human complex emotional states. Soft computing techniques are applied in mapping imprecise input into emotions. The fuzzification strategy is proposed which can successfully quantify the extracted information from each modality into a fuzzified value. Both Mamdani-tpye and Sugeno-type system were applied and the results were compared. Although both systems perform well, Sugeno-type system is more suitable for adaptive cases for constructing fuzzy models. This adaptive technique can be used to customize the membership functions so that the fuzzy system best models the data.

Chapter 4

Head Movement Detection

4.1 Introduction

People use the movement direction of their heads to convey emotional and cognitional information. For example, a nodding head usually indicates yes and a shaking head usually indicates no. In order to estimate head pose in computer vision, a thorough review has been made by Murphy-Chutorian et al. [166]. They summarized 8 methods to estimate head pose which are the following: appearance template methods, detector array methods, nonlinear regression methods, manifold embedding methods, flexible models, geometric methods, tracking methods, and hybrid methods. In their survey, the comparisons between each method and their advantages and disadvantages have been provided according to the functional requirements.

The goal of this thesis is to recognize emotion by using the emotional model provided by psychology experts. In this model, head nods and shakes as well as their frequencies are utilized as an information channel from the body language of the head. Complex head pose estimation is not needed. Therefore, we simply apply the tracking methods to estimated head movement direction using a single webcam. The general procedure of head movement detection is shown in Figure 4.1.

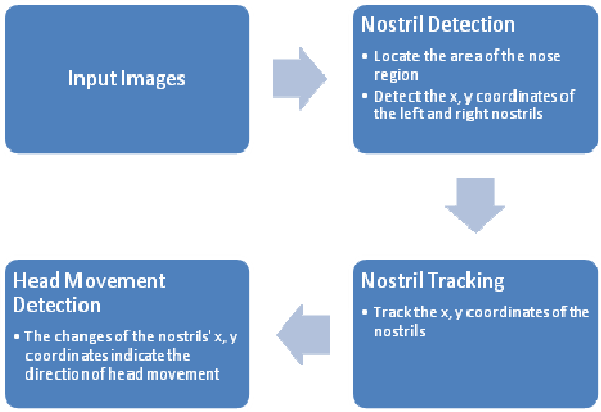


Figure 4.1 General steps of proposed head movement detection

4.2 Nose Region Segmentation

In order to build a system capable of automatically recognizing head movements, the first step is to detect and extract the human face from the background image. This is the first step; however, it is also one of the most important steps in emotion detection, since all the following processes are based on the successful detection of the face. The face detection algorithm should meet the following requirements:

1) Able to extract face image from a clutter background.
2) Able to extract face image from various lighting conditions.
3) Tolerant to a certain degree of head in-plane and out-plane rotation.
4) Expression-independent, that is, can extract face under different facial expressions.
5) Able to extract face regardless of the variation in facial size, color, and texture.

To meet all of the requirements above, we make use of a robust and automated real-time face detection scheme proposed by Viola and Jones [106], which consists of a cascade of classifiers trained by AdaBoost. In their algorithm, the concept of "integral image" is also introduced to compute a rich set of Haar-like features (see Figure 4.2). Each classifier employs the integral image filters, which allows the features be computed very fast at any location and scale. For each stage in the cascade, a subset of features is chosen using a feature selection procedure based on AdaBoost. The Viola-Jones algorithm is approximately 15 times faster than any other previous approaches while achieving equivalent accuracy as the best published results [107]. Figure 4.3 shows the detection of the human face using Viola-Jones algorithm using an image from the JAFFE

database [115].

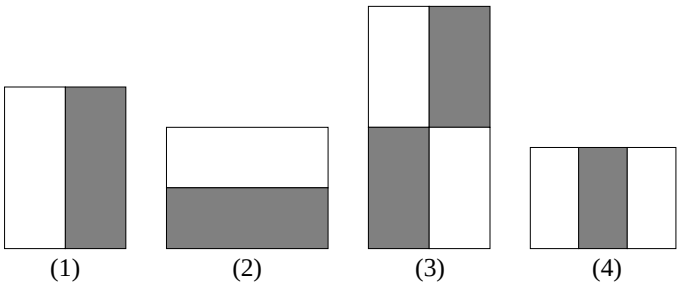
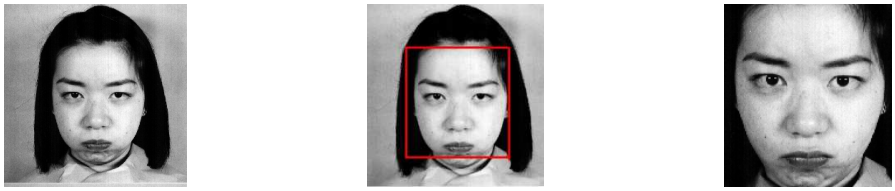


Figure 4.2 Haar-like features



a) Input Image b) Face detection c) Detected face

Figure 4.3 Example of face detection

In the human face, nostrils are relatively static feature points that will not be significantly affected by different facial expressions. Therefore, the coordinates of these points can be used to track and predict head movements. We focused on the nose region instead of the whole face image to detect feature points. There are two advantages of focusing on the nose region: (1) It can eliminate interruption from other facial features, and (2) incorrect detection of nostrils does not affect the recognition of head movement. This is because any facial features within the nose region are relatively static and will not be affected by facial expressions. Tracking and analyzing the motion of any feature in the nose region can also predict head movements.

We obtained the nose region by geometrical proportion of the human face since the position of the nose region is always approximately 2/4 to 3/4 of the height and 2/7 to 5/7 of the width of the face. Let (x_1, y_1) be the coordinates on the upper left corner of the face region, and (x_2, y_2) be the coordinates on the bottom right corner of the face region. We have:

$$X_L = 2*(x_2 - x_1)/7,\ Y_L = 2*(y_2 - y_1)/4 \quad (4.1)$$

$$X_R = 5*(x_2 - x_1)/7,\ Y_R = 3*(y_2 - y_1)/4 \quad (4.2)$$

where X_L and Y_L are the coordinates on the upper left corner of the nose region; X_R and Y_R are the coordinates on the bottom right corner of the nose region. The following steps are all based on the located nose region.

4.3 Head Movement Analysis

4.3.1 Nostril Detection and Tracking

Our nostril detection method was inspired by the methods described by Vukadinovic and Pantic [116]. We have extended their methods by first applying the Harris corner detection algorithm [117] to automatically detect feature point candidates instead of manually labeling them in the training stage. The Harris corner detection algorithm not only can detect corners but also any number of isolated feature points in the region of interest (ROI) simultaneously. The number of feature points is flexible and can be predefined in the program. Since nostrils are obvious feature points in the nose region, 10 interest points are enough for our recognition. The order of the selection of the interest points begins with the one that has the maximum eigenvalue and proceeds consecutively until it reaches the number we defined. Apparently, only two of these 10 feature points are the nostrils. If we define only two interest points in the ROI, in most cases, they would be the nostrils. However, in some cases, such as when a person has a mole in the nose area, there is a high possibility that the mole will be misinterpreted as feature points. Therefore, a nostril classifier needs to be trained in order to accurately detect nostrils. An automatic nostril detection method using the Gabor feature-based boosted classifier [118] was applied.

In the training stage of nostril detection, the 10 interest points that were detected in the previous step were used both as positive samples and negative samples. We manually identified the 2 out of 10 interest points that were positive feature points (nostrils). The rest of the eight detected interest points were used for the negative ones. Gabor filters [119] with eight orientations and six spatial frequencies were applied for feature extraction. The feature vector for each sample point was extracted from a 5*5-pixel bounding box centered at each positive and negative sample point. The 5*5-pixel bounding box was extracted from both the grayscale image and the Gabor filter bank (consisting of 8*6=48 representations). Therefore, 49*5*5 features are used to represent

one feature point. Since there are four negative points and one positive point for each nostril, the size of the training data matrix would be 5*49*5*5, which is computationally expensive. To solve this problem and avoid redundancy, the GentleBoost algorithm [118] was used to reduce the dimensionality.

In the testing stage, a 5*5-pixel sliding bounding box was used to slide across the ROI pixel by pixel. The GentleBoost classifier outputs a response depicting the similarity between the trained feature point model and the current sliding bounding box. When the entire ROI region has been scanned, the position with the highest response shows the location of the feature point (nostril). This method is very effective for feature point detection, especially in the case of person-specific applications since the false candidates, such as the moles are already trained as negative samples and will not be misidentified as true feature points.

We applied the iterative Lucas-Kanade (LK) method with pyramids [120] for real-time tracking of the nostrils. This method implements the sparse iterative version of the LK optical flow in pyramids and calculates the coordinates of the nostrils on the current video frame given their coordinates on the previous frame. Since the head may move back and forth in front of the camera, the size of the face segment changes. Therefore, multi-level matching is desired. The window size for computing the local coherent motion was set as 25*25, and the depth of the pyramid level was three. In extreme cases when the speed at which the object is moving is too fast, the feature points may be lost. For the purpose of detecting the natural rhythm of the shaking and nodding of the head, we found that this tracking algorithm was fairly reliable.

4.3.2 Head Movement Analysis

After the nostrils had been detected and tracked, the coordinates were used to determine the head movements. We adopted a statistical pattern matching approach, which was trained and tested on real data to detect head movement. To improve recognition accuracy and to achieve real-time performance, a boosted-based pattern analyzer [118] was used to adaptively select the best features by using a linear combination of individually weak classifiers and combining them into a strong classifier. Initially, the boosted analyzer assigned equal weight to each training sample. For the next stage, more weight was added to the training samples that were missed in the previous stage. The iteration went on by adding new classifiers until the overall accuracy met the

desired requirement. There was a tradeoff between the overall recognition accuracy and the speed. For each type of head movement (head nods, head shakes and stationary), 20 weak classifiers in total were selected.

To determine head movements, a feature extraction module was used to locate the nostrils and map out the displacement onto the x, y coordinates. When the head movement is vertical, the shift of the nostrils in the y-coordinates is greater than that of the x-coordinates. On the other hand, when the head movement is horizontal, the shift of the nostrils in the x-coordinates is greater than that of the y-coordinates. If the shift in the displacement of the nostril is within a user-defined limit on the x and y axes, then the head movement is considered to be still. Suppose the coordinate of a nostril is (x_{n-1}, y_{n-1}) and (x_n, y_n) in two consecutive frames respectively, then $| y_n - y_{n-1}| >> | x_n - x_{n-1}|$ indicates head nods, and $| y_n - y_{n-1}| << | x_n - x_{n-1}|$ indicates head shakes. If both $| y_n - y_{n-1}|$ and $| x_n - x_{n-1}|$ are lower than a certain threshold, then stationary status has occurred. For each frame, an observation sequence *O* consisting of 10 consecutive frames is generated:

$$O = \{[(x_n - x_{n-1}), \dots , (x_{n-9} - x_{n-10})], [(y_n - y_{n-1}), \dots , (y_{n-9} - y_{n-10})]\} \quad (4.3)$$

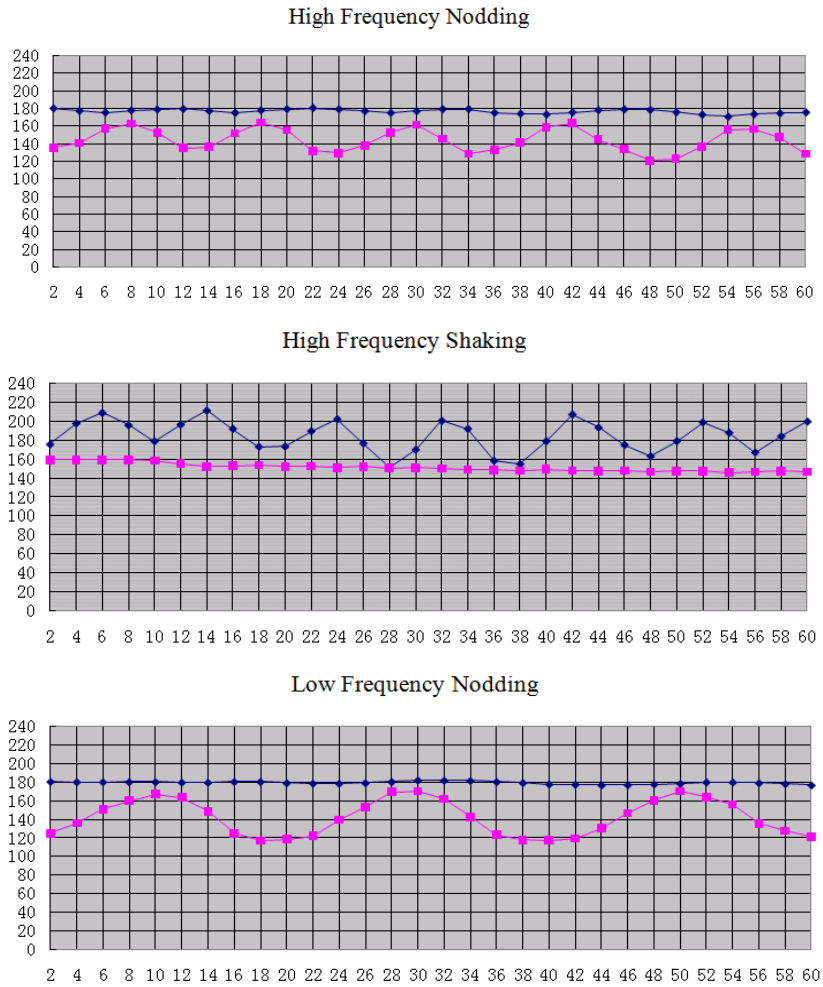
The coordinate differences between each of these frames were used as inputs for the boosted classifier to recognize head nodding and shaking for each frame. The nodding and shaking of the head can have different frequencies and amplitudes depending on the context. Ten is considered to be a reasonable number of an observation sequence, since 1) It can prevent misclassification of a random head movement as a nod or shake, but is enough to retain recognition of subtle head movements. 2) We do not want the time needed to perform a series of nod or shake to be too long (the frame rate is 15 frames/sec). Only when all 10 consecutive frames respond with the same output, is the final interpretation being considered accurate.

The goal of head movement analysis is not only to observe the direction of head movement but also to calculate the frequency of head movement. Head movement frequencies were quantified into values for the further fuzzy inference system inputs. To achieve this, tracking results of the nostrils were used.

Figure 4.4 shows examples of nostril tracking results under five different conditions: high frequency nodding, low frequency nodding, high frequency shaking, low frequency shaking and stationary. The blue plots indicate the x-coordinates of the nostril points and the pink plots indicate the y-coordinates of the nostril points. The x-axes represent the

number of detected nostril points and the y- axes represent the corresponding pixel value of these points in the 320*240 images.

To determine the frequency of head movement, we applied the Fast Fourier Transform (FFT) [121] to transform the tracking results from the time domain to the frequency domain. Tracking results of only one nostril is needed for the frequency analysis. The head movement direction is already known in the previous step. If the head movement direction is nodding, the y coordinates of nostril tracking results is applied by FFT. On the other hand, if the head movement direction is shaking, the x coordinates of nostril tracking results is applied by FFT. After FFT, the highest peak in the spectrum shows the main frequency for the input signals. Figure 4.5 shows an example of the result after FFT for the status of Low-Frequency-Nodding. For each video sample, we recorded the coordinate of the nostril tracking result and calculated the frequency value using FFT. These frequency values were used to quantify the input variables we defined in the Emotion Recognition chapter.



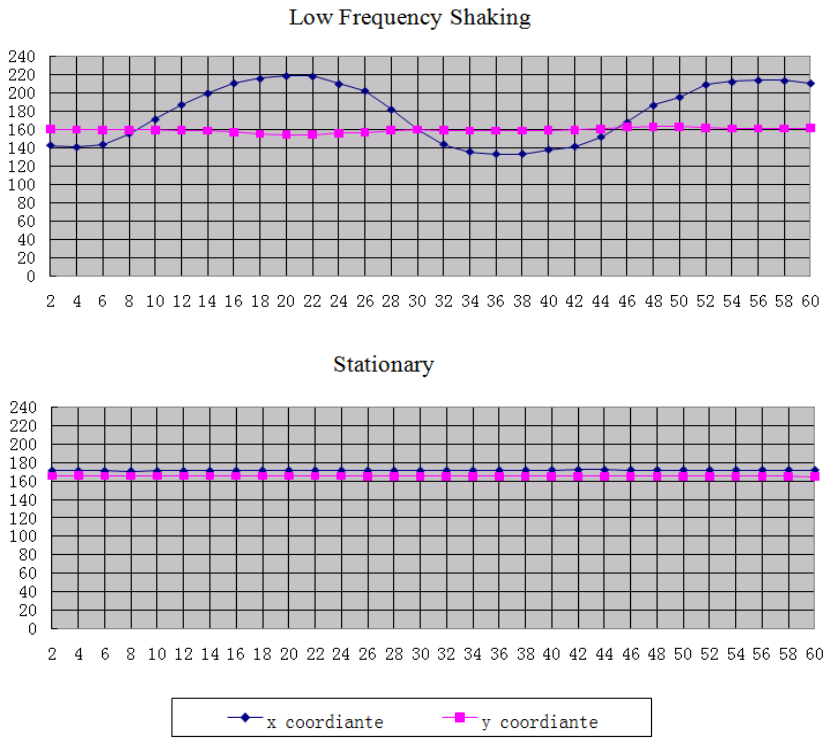


Figure 4.4 Example of nostril tracking results

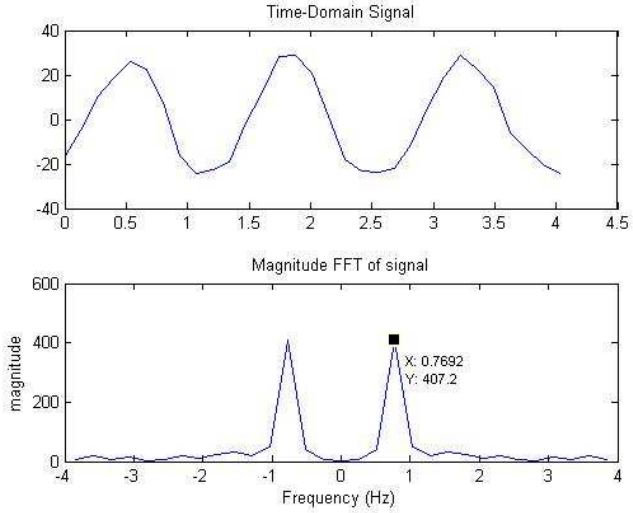


Figure 4.5 Example of FFT for low frequency nodding

4.4 Experiment and Results

4.4.1 Nostril Detection Evaluation

To evaluate the performance of the proposed method, we tested our nostril detection method on both on the JAFFE database [115] and images collected from a webcam. The original purpose of JAFFE database is built for extracting facial feature information according to the six basic facial expressions and neutral expression. There are 10 Japanese female in the database, each posing three or four examples of each of the six basic facial expressions and neutral face for a total of 219 images of facial expressions. All of the images in JAFFE database are frontal face images and each of the images is labeled with a defined expression. Figure 4.6 shows the examples of images from the JAFFE database. From the left to the right, the corresponding facial expressions are angry, disgust, fear, happiness, neutral, sadness, and surprise.

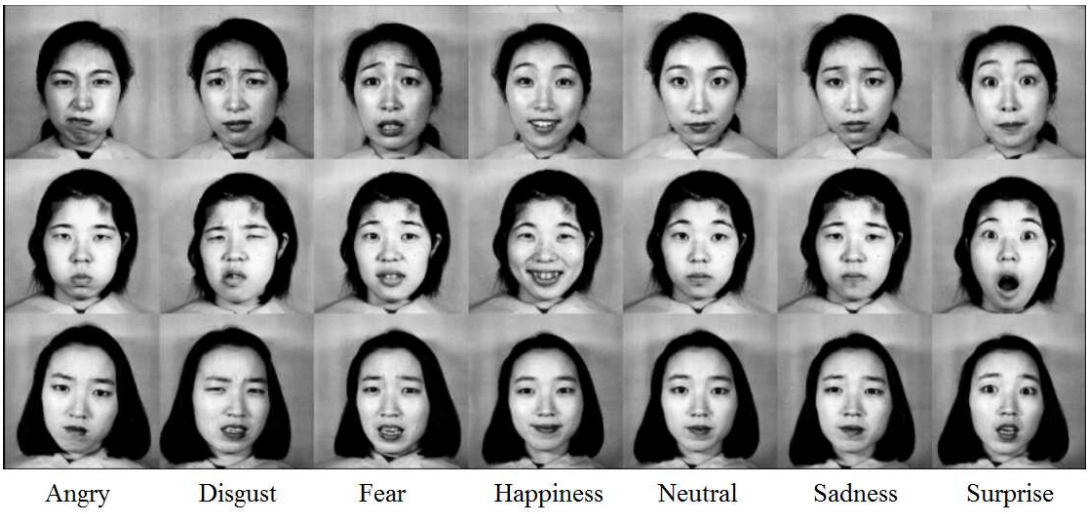


Figure 4.6 Examples of the JAFFE face images

The images collected from the webcam were taken by two volunteer colleagues, one male and one female. A total of 200 image samples were collected using the Microsoft LifeCam 1.4 with a resolution of 640*480. In order to complicate the data, the participants were asked to perform different facial expressions. Some of the collected images have in-plane and out-plane head rotations. For both of these two experiments, each automatically detected nostril point was compared to the manually annotated nostril point (considered as true point). If the automatically detected feature point was placed

within 3 pixels distance from the manually annotated true point, then this point is regarded as successful detection. Figure 4.7 and Figure 4.8 show some examples of nostril detection results using the JAFFE database and webcam respectively. The detection accuracy of JAFFE database is 93%. No images from the database were missed. The measurements of the mean and variance of the distance between a detected point and true point of the seven facial expressions from the JAFFE database are listed in Table 4.1. The nostril detection accuracy using webcam data is 95% and 11 images were missed. Our nostril detection is based on the detected human facial images via algorithm proposed by Viola and Jones [106], which is sensitive to head in-plane and out-of-plane rotation. Therefore, the nostril detection method will fail when the images have excessive head in-plane and out-of-plane rotation invariance over ±15°.

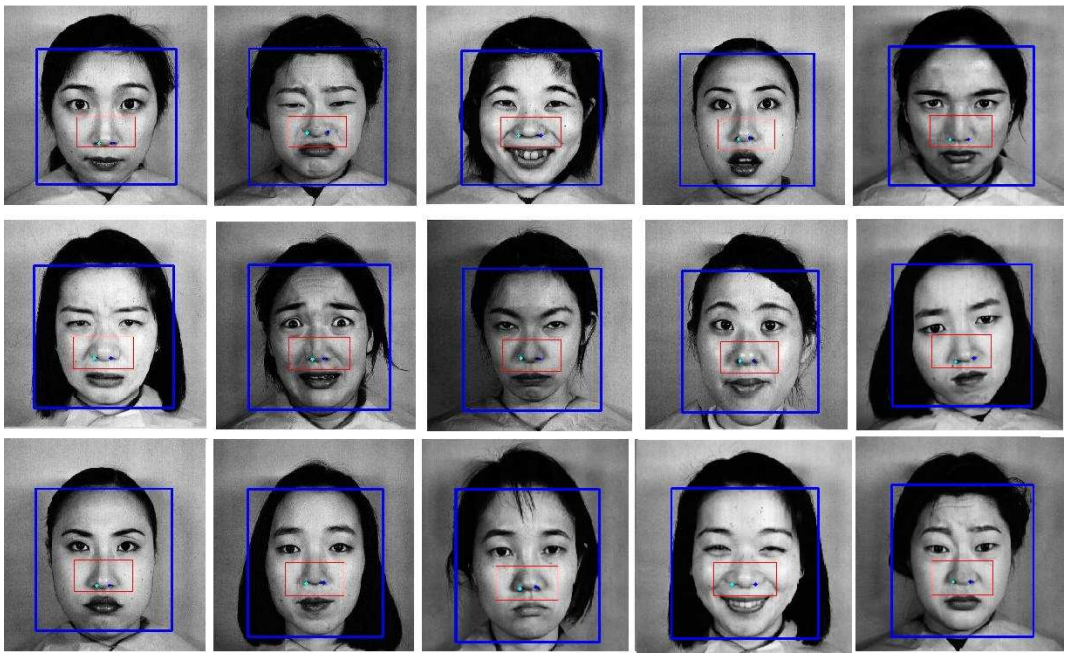


Figure 4.7 Examples of nostril detection results using the JAFFE database

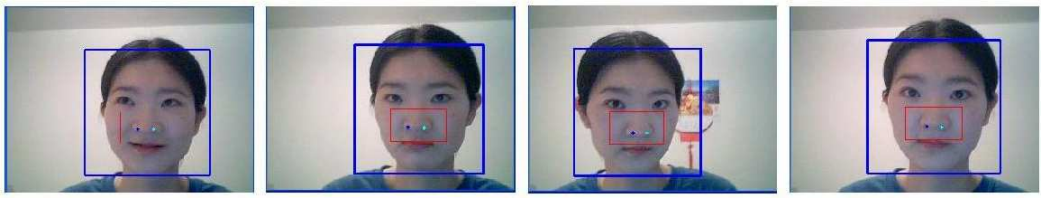


Figure 4.8 Examples of nostril detection results using webcam images

Table 4.1 Measurements of the distance between detected point and true Point of the seven facial expressions from the JAFFE database

Emotion	*Angry*	*Disgust*	*Fear*	*Happy*	*Neutral*	*Sad*	*Surprise*	*Overall*
Mean	1.048016	1.452378	1.626345	1.712985	1.464439	1.47756	1.406209	1.45388
Variance	1.01864	1.450082	0.868573	1.214302	0.802119	1.190863	0.672096	1.05854

4.4.2 Head Movement Detection Evaluation

In order to evaluate the head movement detection performance, four subjects (two male and two female) participated in the experiment. The Microsoft LifeCam 1.4 at a resolution of 320*240 and 15 frames/s was used for the head movement analysis experiment. A database was collected of natural head nods and head shakes by asking a number of questions that the subjects were instructed to answer with a head nod, a head shake or keeping his or her head stationary. Examples of such questions are listed in Table 4.2. For the questions that required the participants to make a choice (for example, question 7 in Table 4.2), they were asked to answer the question by keeping their head stationary. A total of 100 samples were collected by asking 25 questions from each of the participants. The participants were asked to sit with frontal face and perform facial expressions while answering the questions. Data was collected during different times of the day near the window in order to vary the lighting conditions. We randomly selected 40% of each of the head nods, shakes and stationary samples for training. To train a head nod classifier, head nod samples are used as positive samples, and the rest of the head shake and stationary samples are used as negative samples. The same strategy was applied for head shakes and stationary classifiers, in turn. In the recognition stage, the classification decision was determined by the classifier with the largest value. The recognition results are shown in Table 4.3. The errors could be caused by in-plane and out-plane rotations of the head movement.

Table 4.2 Example questions

	Questions
1	Are you a university student?
2	Are you male?
3	Are you female?
4	Do you like basketball game?
5	Are you Canadian?
6	Do you like travelling?
7	Do you like apples or oranges?

Table 4.3 Head movement test results

Head Movements	*Hits*	*Missed*	*False*	*Recognition Rate*
Nods	40	0	2	95%
Shakes	30	0	2	93.3%
Stationary	30	0	1	96.7%

4.5 Summary

Nodding and shaking are very common head gestures that can send and receive messages and express emotions and cognitions. The frequency of head movement (high frequency movement or low frequency movement) is taken into consideration as well as head nods and head shakes in the process of emotion context recognition. We proposed a simple, fast, and effective automatic head movement detection approach by simply using a webcam. The movement direction of the nostril points are used to estimated the head movement direction. The head movement frequency is acquired by analyzing the tracking results of the coordinates from the detected nostril points. Five head movement status can be detected for future emotion estimation including high frequency head nodding, low frequency head nodding, still, low frequency head shaking, and high frequency head shaking.

Chapter 5

Eye Gaze Analysis

5.1 Introduction

Eye gaze information is neglected in the past research of emotion recognition, which has been proofed by psychologists to be an important factor during emotion recognition. Eye contact plays an important role in human-human interactions. It is the key for establishing and maintaining conversations with others. Human eyes can provide the most apparent and direct cues that form the impressions of a counterpart [122]. One can often anticipate the feeling and thought of the opposite side. For example, nervousness can be noticed by observing a person's eye contact, perspiration and stiffness [123]. In face to face communication, lack of eye contact is sometimes considered to be rude or inattentive [123].

Current eye gaze detection systems are mainly focusing on accurate eye gaze direction estimation, for example, the exact position a person is looking at. This needs the precise detection of black or white pupil detection and eye tracking from very high quality images from different angles [166]-[177].

In order to achieve this, complex algorithms and devices (e.g. multiple cameras, infrared camera, or sensors) is necessary. In our thesis, we are analyzing eye gaze from

the emotion point of view. Only two status of eye gaze is needed for the emotion model. Therefore, we utilize the geometrical relationship of human organs between nostrils and two pupils to achieve this task.

5.2 Eye Detection

The purpose of this chapter is to determine whether the subject's gaze direction was direct or averted. A geometric relationship of human face organs was used to analyze eye gaze status. Since there were only two statuses that we considered, direct gaze and averted gaze, we compared the face image to a known direct gaze image as the template in order to evaluate the possible gaze status. The positions of nostrils and pupil locations were analyzed as evaluators. We defined the geometry of the eyes and nostrils as shown in Figure 5.1.

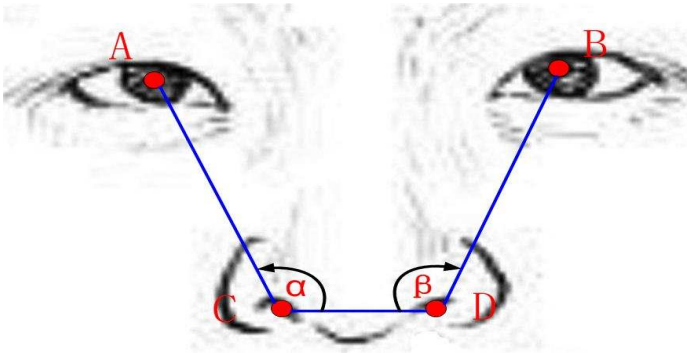


Figure 5.1 The geometrical eye and nostril model

We computed the four parameters r_R, r_L, α, and β where r_R, r_L are defined as follows:

$$r_R = \frac{|AC|}{|CD|}, \quad r_L = \frac{|BD|}{|CD|} \tag{5.1}$$

α and β are angles in radians. We denoted the values of the four parameters of the template in which the subject's gaze status was a direct gaze as r_{R0}, r_{L0}, α_0, and β_0. For each face image, we defined an evaluation parameter S, computed as follows:

$$S = |r_R - r_{R0}| + |r_L - r_{L0}| + |\alpha - \alpha_0| + |\beta - \beta_0| \tag{5.2}$$

The nostrils are relatively static points in frontal face images. The pupils move when eye gaze status changes. Assuming the face images are all the same size, the length and orientation of CD are stable. When the subject is in averted gaze status (i.e., he or she

looks left, right, up, or down), |AC| and |BD| change, and the angles α and β change as well. To eliminate the effects of image size, we used the fractions of lengths r_R and r_L instead of absolute values |AC| and |BD|. In the case of direct gaze status, $|r_R - r_{R0}|$, $|r_L - r_{L0}|$, $|\alpha - \alpha_0|$, and $|\beta - \beta_0|$ are approaching zero, respectively. On the contrary, these values are greater than zero in the averted gaze status. By adding these values together, we can distinguish the direct gaze and averted gaze.

The average *S* value was computed for each test video and normalized into the range of 0 to 1. The final eye gaze input value of the fuzzy inference system was 1-*S*.

5.3 Pupil Detection

We used a knowledge-based method to detect the location of the pupil. To ensure accurate pupil detection, our approach was built on the following conditions:

1) The frontal view face images, which were picked out from the sequence of video frames, were used.
2) Only the image segment of the human face, which is detected from the frontal view face image, was analysed.
3) The environment light uniformly illuminated the whole face.

5.3.1 Detection of Frontal View Face Image

In order to keep the consistency of the geometric relationship of face organs such as eyebrows, eyes, nose and mouth, we considered only frontal view face images in which the subjects were facing the camera directly. A frontal view image was used as a reference. As stated in Chapter 4, the coordinates of nostrils are determined for each frame. There are two bounding boxes that are defined according to the location of the reference. Let x^0, y^0 be the reference nostril coordinates. Two bounding boxes corresponding to the left and right nostrils are defined as follows:

$$B_{nostril} = \{(x, y) \,|\, x \in [x^0 - \varepsilon, x^0 + \varepsilon],\ y \in [y^0 - \varepsilon, y^0 + \varepsilon]\} \tag{5.3}$$

where ε is the small number that we define as a threshold. Let (x, y) be the detected location of the nostrils. If the two nostrils are both inside the bounding boxes， then this frame is a frontal view image.

5.3.2 Generation of the Face Mask

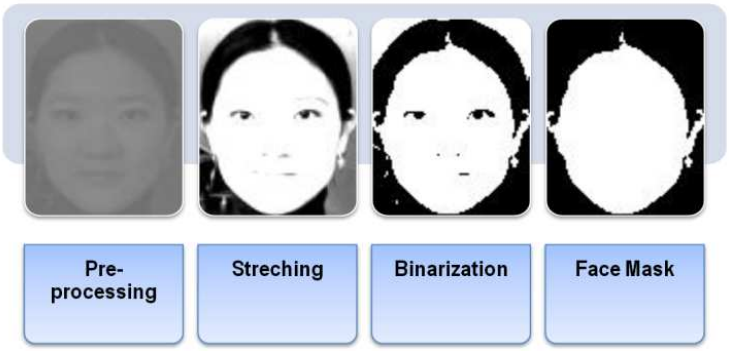


Figure 5.2 Generation of the face mask

A face mask is used for removing the information that does not belong to the face region. The process for generating a face mask is shown in Figure 5.2. In the first step, the color image is converted into gray-scale. To unify the brightness intensity distribution of the image, the brightness of the image is normalized using the following equation:

$$F_n(x,y) = Mean_0 + sign \times \sqrt{Var_0 (F(x,y) - Mean)^2 / Var} \tag{5.4}$$

where $F(x, y)$ and $F_n(x, y)$ are the input face image and normalized image, respectively. $Mean_0$ and Var_0 are the mean and variance of the destination image. *Mean* and *Var* are the mean and variance of input image; *sign* is equal to 1 when $F(x, y) > Mean$, and -1 when $F(x, y) <= Mean$.

Secondly, enhancing the contrast of the image improves the performance of binarization. Thus, we applied gray-scale stretching [124] to increase the contrast as shown in the following equations:

$$F_{str}(x,y) = \begin{cases} 255 \times \dfrac{F_n(x,y) - low}{(high - low)} & , low \le F_n(x,y) \le high \\ 255 & , \quad F_n(x,y) > high \\ 0 & , \quad F_n(x,y) < low \end{cases} \tag{5.5}$$

where *low* and *high* are set to $Mean_0 \cdot P_0\%$ and $Mean_0 \cdot P_1\%$, respectively.

For binarization, we assumed that the hair is darker than the face skin. Consequently, we simply took the mean value of the stretched image as the threshold to binarize the image. The threshold can be adjusted for opposite cases when the skin is darker than the hair. As shown in step three of Figure 5.2, the face skin area turns white.

In the last step, we want to further remove the face organs such as the eyes, lips, and nostrils to turn the entire face region white. For most cases, the face skin is brighter than the face organs; the eyes, nostrils and lips appear to be black. To take out these small black areas, we find all connected components in which all pixel values are the same and connected to each other from the binary face image. Later on, the pixels of these components are counted respectively, and then the small black areas, whose pixel number was less than a certain percentage of the total pixels in the image, are turned white. Using the same technology, we remove the small white areas in a large black region for the purpose of turning the white pixels in the background to black. The resulting image is used as the face mask. The face mask is a binary matrix that has the same size as the original face image.

5.3.3 Determination of Eye Location

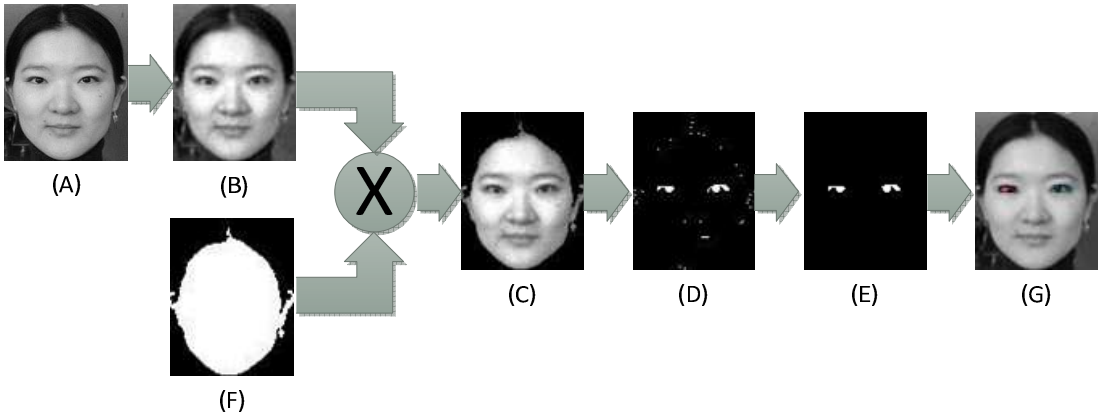


Figure 5.3 Determination of eye location

The procedure for locating eyes is illustrated in Figure 5.3. First, we apply histogram equalization to the original gray face image ((A) of Figure 5.3). The resulting image of histogram equalization ((B) of Figure 5.3) is denoted by $F_{hsq}(x, y)$. We multiply the binary face mask ((F) of Figure 5.3) obtained from Section 5.3.2 with $F_{hsq}(x, y)$ for the purpose of removing the confusing objects not located in the face area, such as the hair region. The image obtained after multiplying the face mask is shown in (C) of Figure 5.3.

The following binarization equations are applied to the masked face image $F_{mf}(x, y)$ to extract the dark components on the face:

$$F_b(x,y) = \begin{cases} 1, F_{mf}(x,y) < 95 \\ 0, \quad otherwise \end{cases} \tag{5.6}$$

Based on our test, we found the pixel number of the eye region to be less than 3% percent and greater than 0.1% of the total pixel number. We removed the connected white component that was larger than 3% and smaller than 0.1% of the total image area. The resulting image, in which the white connected pixels are extracted face organs, is shown in (E) of Figure 5.3.

Among the extracted face organs, eyebrows, lips, nostrils, and other noise regions are the false candidates. The knowledge of the face's geometry and spatial relationship is applied to locate the eyes. The proposed algorithm is described as follows:

Given the binary face object image $F_{bfo}(x,y)$, horizontal integral is computed.

$$H(y) = \sum_{x=1}^{W} F_{bfo}(x,y) \tag{5.7}$$

1. After smoothing the $H(y)$, local peaks are detected. These peaks are taken as the possible y-coordinates of face objects. *Y-coordinate* of first local peak above 40% of the image height from the bottom is taken to locate eyes. Let y_{eye} denotes the peak index. As illustrated in Figure 5.4, the left panels are the horizontal integrals; red asterisks denote the detected peaks and the right panels illustrate the labelled face objects.

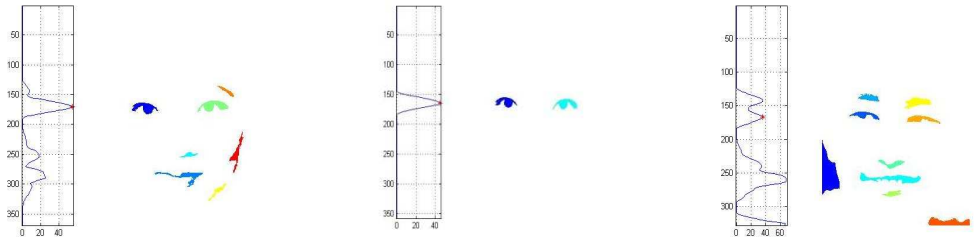
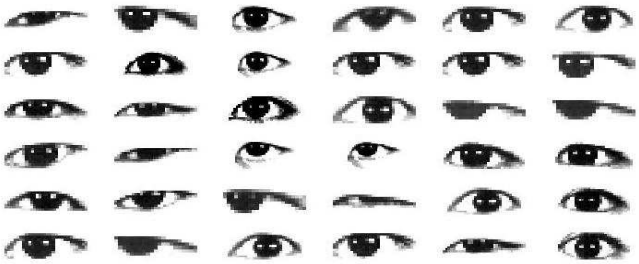


Figure 5.4 The examples of applying horizontal integral to determine y-coordinate of eye location

2. The white areas along the horizontal line y = y_{eye} are candidates of eyes. As we know the subject face is frontal orientated and in the center of the image, the objects next to the left and right of the x = $W/2$ (where W stands for the width of the face image) are two eyes.

3. Two bounding boxes, which are confined by the minimal and maximal indices of detected objects in step 2, are applied to extract right and left eye images. Figure 5.5 shows some examples of detected eye images.

(a) Right eyes



(b) Left eyes

Figure 5.5 The examples of detected eye region segments (from JAFFE database)

5.3.4 Determination of Pupil Location

The detection of pupil locations is broken into two steps: (1) estimating the *y*-coordinates of two eyes and (2) estimating the *x*-coordinates of two eyes. The local horizontal integral projections of two grayscale eye regions are computed separately. The horizontal integral is defined by:

$$H(y) = \sum_{x} F_{hsq}(x, y), \quad (x, y) \in R_{eye} \tag{5.8}$$

where R_{eye} represents the region of the eye determined in the section 5.3.3.

Because the darkest area is most likely the centre of the iris, the y-value corresponding to the minimal $H(y)$ value is selected as the y-coordinate of the pupil. The vertical integral projections are computed as follows:

$$V(x) = \sum_{y=y_p-W_p}^{y_p+W_p} F_{hsq}(x, y), \quad (x, y) \in R_{eye} \tag{5.9}$$

where y_p denotes the detected y-coordinate of the pupil, and W_p is the estimated pupil diameter. The x-value corresponding to the minimal $V(x)$ value is selected as x-coordinate of the pupil.

5.4 Experiment and Results

5.4.1 Pupil Detection Evaluation

The JAFFE database of frontal face images was chosen for evaluation of the pupil detection procedure. The resolution of each image is about 256*256 pixels. We chose $Mean_0$=127, Var_0=85, P_0=80, P_1=100 (where P_0 and P_1 are the minimum and maximum percentages of pixels from the mean of the histogram). The detected pupils are shown in Figure 5.6. Since the face detection algorithm is a scale invariant [107], the size of the detected face region will not be a constant. The eye regions were detected from these face regions with various sizes based on geometrical characteristics of the human face. Therefore, the size of the face image does not affect eye detection. In order to further test the performance of the method, we took the images collected from the webcam (named WebCam-1) in Chapter 4.4.1 for nostril detection. We decreased the size of the same set of images to 320*240 and named them WebCam-2. The size of the detected face region varies and ranges from 274*274 to 194*194. The comparison of the pupil detection results is shown in Table 5.1. If the detected point was placed within 3 pixels distance from the manually annotated true pupil, then this point is regarded as successful detection. The results show that this algorithm is not sensitive to the scale of the image. However, when the image quality is poor and the eyes are no longer noticeable, this algorithm will fail. The reasons for the failure of detection are as follows: (1) The environmental light does not uniformly illuminate the face; for example, one side of the face is too dark to detect the eye location. (2) The head does not face the frontal plane; thus one eye is incorrectly taken as part of the face mask. (3) The head is tilted too much and two eyes are not aligned horizontally, causing the algorithm to fail. In further tests, the Gaussian noise was added into the image with SNR around 20dB, and there was no significant change of detection rate.

5.4.2 Gaze Direction Evaluation

To validate the proposed eye gaze detection method, we analyzed images captured by the webcam at a resolution of 640*480. Five colleagues volunteered to participate in the experiment. Each participant was asked to repeat all the possible combinations of seven facial expressions with five different eye gaze directions (direct, up, down, left and right) five times. Therefore, in total, 5*25*7=875 images (175 for direct gaze and 700 for avert gaze) were collected. Table 5.2 shows an example of eye gaze evaluation data by comparing five images with the reference image. In this table, columns "r_R", "r_L", "α", "β", and "*S*" are the five parameters defined in formula (5). The column "*Result*" represents the detection result for eye gaze. We set a threshold of 0.25 for "*S*". When the values of "*S*" are less than the threshold, we categorized the image into direct gaze using 1 to represent it; it was otherwise categorized as an averted gaze using 0 to represent it. The value of *S* was then normalized into the range of 0 to 1. For each subject, we calculated his or her maximal *S* value based on these collected images with respect to the reference image of direct gaze. The average value of these maximal *S* values was used for normalization. The column "*Normalization*" shows the result after normalization and column the "*Fuzzy Input*" column shows the result of 1 minus the normalized value. The final value will be the quantified eye gaze value of fuzzy input for emotion detection. Table 5.3 shows the measurements of the mean and variance value of the parameter *S* from the collected images under different gaze directions. Since direct gaze images should compute the smallest *S* values when compared to the reference image, the mean value of "*Direct*" is the smallest among all gaze directions.

Figure 5.6 Example of pupil detection results

Table 5.1 Comparison of pupil detection results

	Original Image Size	*Detected Face Region (Approximately)*	*Recognition Rate*
JAFFE	256*256	173*173	96.7%
WebCam-1	640*480	274*274	97%
WebCam-2	320*240	132*132	96.5%

Table 5.2 Example of eye gaze analysis

	r_R	r_L	α	β	S	*Result*	*Normalization*	*Fuzzy Input*
Reference	2.2209	2.2224	2.0700	2.0081	0.0000	1	0.0000	1.0000
Direct	2.2288	2.2566	2.0670	1.9952	0.0579	1	0.0986	0.9014
Up	2.3707	2.5228	1.9745	2.0149	0.5524	0	0.9415	0.0585
Down	1.9444	2.0142	2.0543	2.0348	0.5273	0	0.8986	0.1014
Right	2.0587	2.3331	1.9270	2.0798	0.4877	0	0.8311	0.1689
Left	2.2306	2.0130	2.1663	1.8588	0.4647	0	0.7920	0.2080

Table 5.3 Measurements of different gaze direction

Gaze Direction	*Direct*	*Up*	*Down*	*Left*	*Right*
Mean	0.09231	0.52048	0.4821	0.4759	0.49456
Variance	0.004012	0.004217	0.010959	0.00266	0.002337

5.5 Summary

The goal of gaze status study is to analyze the subject's gaze direction (direct gaze or averted gaze). Our approach is knowledge based and preliminary results are taken as assumptions. After we detect the locations of the nostrils and the pupils, the geometrical relations between these four points are studied and evaluated into a parameter in order to represent the gaze status. The quantified attributes from all the images for eye gaze detection are then used as fuzzy inputs to be mapped into emotional states.

Chapter 6

Facial Expression Recognition

6.1 Introduction

The objective of this chapter is to recognize facial expressions that are used in the proposed emotional model. Although only three types of facial expressions are used for emotion detection, the performance of the algorithms is tested on the JAFFE database, which including six basic facial expressions and neutral expression. The whole recognition process is shown in Figure 6.1.

6.2 Multi-step Integral Projection for Facial Image Segmentation

The first step of head movement recognition is to detect the face region of the input image. In order to automatically capture the human face, we adopted a fast and robust face detection algorithm proposed by Viola and Jones [107] using Haar-like feature-based AdaBoost classifiers.

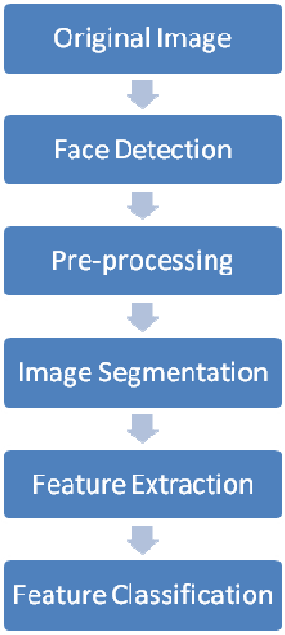


Figure 6.1 The general steps of facial expression recognition

The aim of pre-processing is to eliminate the differences between input images as much as possible; so that we could extract features under the same conditions. Since input images are affected by the type of camera, illumination conditions and so on and so forth, we need to normalize the face images before feature extraction. To combat the effect of these situations, the steps of pre-processing are:

1) Transform the face video into face images.
2) Convert the input color images into gray-scale images.
4) Perform grayscale equalization to reduce the influence due to illumination variety and ethnicity.

To avoid the influence of personal differences, instead of extracting features from the entire face, we attempt to focus on face regions that we are interested in. This can refine useful information and avoid the interference of useless information such as the moles on the face. The most important regions in human faces for classifying expression are eyes, eyebrows and mouth [10]. In this section, we apply the multi-step integral projection method for the segmentation of the facial image.

The basic idea of gray-scale integral projection is to accumulate the sum of vertical gray-scale value and horizontal gray-scale value of an image. The vertical integral projection shows the variations of gray-scale value on the x coordinate of an image. The horizontal integral projection shows the variations of gray-scale value on the y coordinate

of an image. Suppose *I(x, y)* is a gray value of an image, the horizontal integral projection in intervals [y_1, y_2] and the vertical projection in intervals [x_1, x_2] can be defined respectively as *H(y)* and *V(x)*, thus we have:

$$H(y) = \frac{1}{x_2 - x_1} \sum_{x=x_1}^{x_2} I(x, y) \tag{6.1}$$

$$V(x) = \frac{1}{y_2 - y_1} \sum_{y=y_1}^{y_2} I(x, y) \tag{6.2}$$

Figure 6.2 shows an example of vertical and horizontal projection results of an image in JAFFE database.

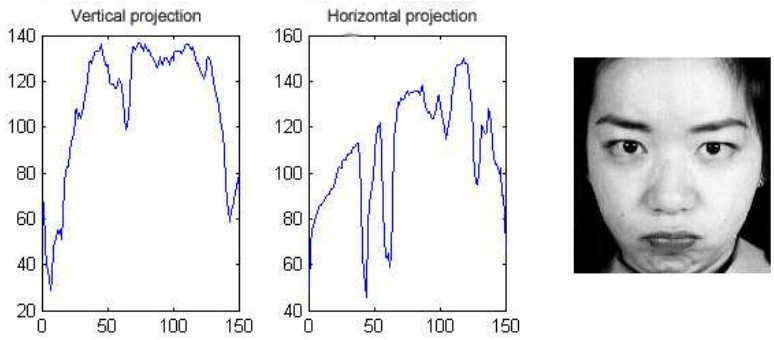


Figure 6.2 Example of integral projection results

In horizontal integral projection, by observing the local minimum, the y coordinate of the center of the eyebrow, eyes, and mouth could be located. Figure 6.3 shows the location of the eyebrows, eyes and mouth by applying horizontal integral projection. From the top to the bottom, the first local minimum value represents the position of the eyebrow; the second local minimum value represents the position of the eyes. From the bottom to the top, the first local minimum value represents the position of the mouth.

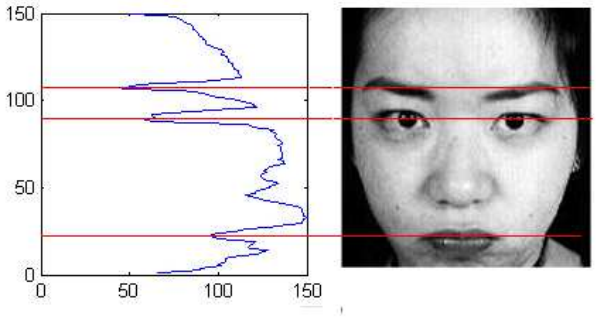


Figure 6.3 Eyebrows, eyes and mouth location

However, due to image complexity, there might be some small local minimum in the projection curve, which interfere the projection results. Therefore, we need to smooth the integral projection curves by filtering local minimum and local maximum to eliminate disturbing information. For any four points: $A(x_A, y_A)$, $B(x_B, y_B)$, $C(x_C, y_C)$, $D(x_D, y_D)$, it starts at $A(x_A, y_A)$ and ends at $D(x_D, y_D)$, the so-called end points. $C(x_C, y_C)$, $D(x_D, y_D)$ are called the control points. Therefore, any coordinate (x_t, y_t) in a curve is:

$$x_t = x_A \cdot t^3 + x_B \cdot (1-t)^3 + 3 \cdot x_C \cdot (1-t) \cdot t^2 + 3 \cdot x_D \cdot t \cdot (1-t)^2 \quad (6.3)$$

$$y_t = y_A \cdot t^3 + y_B \cdot (1-t)^3 + 3 \cdot y_C \cdot (1-t) \cdot t^2 + 3 \cdot y_D \cdot t \cdot (1-t)^2 \quad (6.4)$$

After smoothening, the enhanced curve and the corresponding y coordinates of the eyebrows, eyes, and mouth are shown in Figure 6.4.

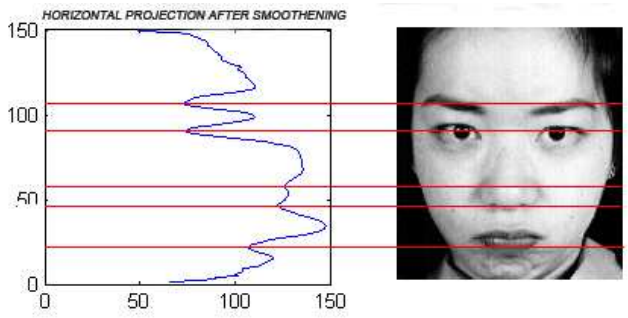


Figure 6.4 Horizontal projections after smoothening

The steps of segmenting the upper and lower y coordinates of eyebrows and eyes areas are as follow: Suppose the vertical length of the face image is H, we define the first local minimum from the top after 0.15H that corresponds to the y coordinate of the eyebrows as H1 and the second local minimum that corresponds to the y coordinate of the eyes as H2 (see Figure 6.5). We would like to segment the eyebrows and eyes area into a bounding box. The upper y coordinate of the bounding box is H1-0.2×VH, where VH=H2-H1; while the lower y coordinate of the bounding box is H2+0.8×VH. All the images in the JAFFE database is tested to defined the threshold in the above steps. For the segmentation, the range of the bounding box is based on the image that has the maximum eyebrows and eyes area in the JAFFE database. This can ensure the bounding box we applied can cover all eyebrows and eyes areas in the JAFFE database for future feature extraction step. In the following steps, the same strategy is used to define the

range of the bounding boxes that use to segment the mouth area.

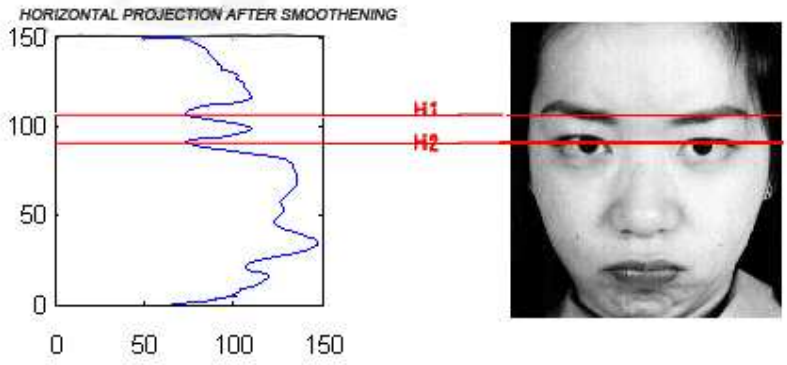


Figure 6.5 The eyebrows and eyes location

The steps of segmenting the upper and lower y coordinates of mouth area are as follow: Suppose the vertical length of the face image is H, we define the first local minimum from the top after 0.7H that corresponds to the y coordinate of the mouth as H4 and the closet local minimum above H4 as H3 (see Figure 6.6). The mouth area is also segmented into a bounding box. The upper y coordinate of the bounding box is H3+0.4×VH, where VH=H4-H3; while the lower y coordinate of the bounding box is H4+0.7×VH.

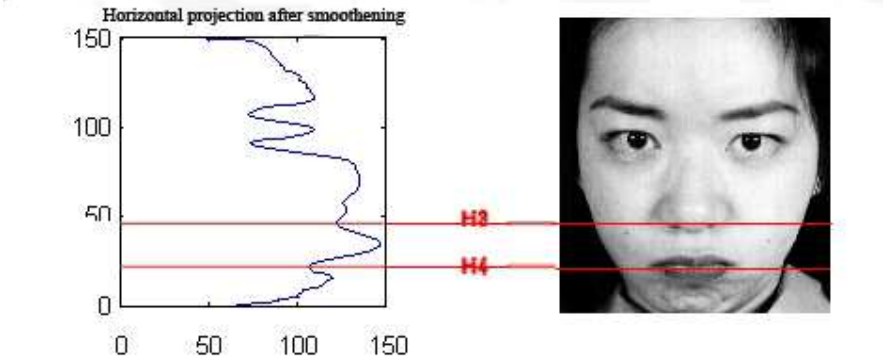


Figure 6.6 The mouth location

After segmenting the upper and lower y coordinates of eyebrow, eyes, and mouth area, the vertical integral projection is then applied for the segmenting of the left and right x coordinates of the eyebrow, eyes, and mouth area. The steps of segmenting the left and right x coordinates of eyebrows and eyes area are as follow: Suppose the vertical length of the image is H, we define the first local maximum from the left to the right greater then 0.75H that corresponds to the left x coordinate of the eyebrows and eyes area

as W1 and the first local maximum from the right to the left greater then 0.75H that corresponds to the right x coordinate of the eyebrows and eyes area as W2. The eyebrows and eyes area is segmented into a smaller bounding box, which W1 is the left x coordinate of the bounding box and W2 is the right x coordinate of the bounding box.

The steps of segmenting the left and right x coordinates of mouth area are as follow: Suppose the vertical length of the image is H. We split the image from the middle into half. The first local maximum from the middle to the left greater then 0.75H that corresponds to the left x coordinate of the mouth area as W1 and the first local maximum from the middle to the right greater then 0.75H that corresponds to the right x coordinate of the mouth area as W2. The mouth area is segmented into a smaller bounding box, which W1 is the left x coordinate of the bounding box and W2 is the right x coordinate of the bounding box.

From the above steps, the bounding boxes of the eyebrows, eyes, and mouth areas can be decided. These segmented areas contain important information for facial expression recognition.

6.3 Facial Feature Extraction based on Gabor Transformation

Gabor wavelets are now being used widely in various computer vision applications due to its robustness against local distortions caused by illumination brightness [108]. They are used to extract the appearance changes as a set of multi-scale and multi-orientation coefficients. Comparing to traditional Fourier transformation, Gabor wavelets can easily adjust the spatial and frequency properties to extract facial features in order to analyze the results in different granularity.

A Gabor wavelet $\psi_{\mu,v}(z)$ is defined as

$$\psi_{\mu,v}(z) = \frac{k_{\mu,v}^{\ 2}}{\sigma^2}\exp\left(-\frac{k_{\mu,v}^{\ 2}z^2}{2\sigma^2}\right)\bullet\left[\exp(ik_{\mu,v}\bullet z)-\exp(-\frac{\sigma^2}{2})\right] \tag{6.5}$$

where $z=(x, y)$ is the point with the horizontal coordinate x and the vertical coordinate y, σ is the standard deviation of the Gaussian window in the kernel, which determines the Gaussian window width. There are three parameters of a Gabor kernel: location,

frequency and orientation. Vector $k_{\mu,v}$ stands for the frequency vector and μ, v defines the orientations and the scale of the Gabor kernel,

$$k_{\mu,v} = k_v e^{i\phi_u} \tag{6.6}$$

where $k_v = k_{\max} / f^v$, $f = \sqrt{2}$, $\phi_\mu = \mu\pi / n$ if n different orientations are chosen, while $k_{\max}$ is the maximum frequency and f^v is the special frequency between kernels in the frequency domain. In our approach, we choose μ=0,1,…,5 and v=0,1,2 therefore, we are using a Gabor filter family with 6*3=18 different orientations and frequencies. In the frequency domain, Gabor filter is an oriented Gaussian with orientation centered at a certain frequency. For all central frequencies, Gabor filter has the response at 0 frequency (a value close to 0). This ensures that Gabor filter is not sensitive to illumination variations.

Gabor representation of an image is the convolution of the image with a family of Gabor kernels. Given an image *I(z)*, the Gabor transformation of *I(z)* can be defined by its convolution with Gabor kernels:

$$G_{\mu,v} = I(z) \bullet \psi_{\mu,v}(z) \tag{6.7}$$

Based on multi-step integral projection and 2D Gabor transformation, we can extract the facial expression features effectively from facial expression images. Several Gabor filters consist into a filter bank, which composed of Gabor filters from various frequencies and orientations. We employ a discrete set of 3 different frequencies and 6 orientation transformation, indexing $\varphi = \frac{\pi}{6}, \frac{2\pi}{6}, \frac{3\pi}{6}, \frac{4\pi}{6}, \frac{5\pi}{6}, \pi$, respectively.

Since the full convolution of face images with different Gabor kernels is very costly in the demand of processing time, we divide the segmented bounding box (acquired from Section 3.3) into grids of $n \times n$ pixels. Gabor transformation is applied on each of these grids instead of the whole segmented bounding box. There is a tradeoff between the processing time and the recognition accuracy based on the selection of grid number. If the grid number is too big, the dimension of the extracted feature vector is too high resulting in redundancy of information that will cost in longer processing time. On the other hand, if the grid number is too small, the extracted information is not enough for correct classification of facial expression. Based on the above considerations and repeated testing, when the grid number is 7*7, the overall system performs best.

6.4 Facial Expression Classification using SVMs

Support Vector Machines (SVMs) have become a powerful solution in pattern recognition areas, especially to the classification problems. SVM is a supervised, non parametric learning system that is developed from statistical learning theory [109], [110]. SVM is initially linear classifier designed for solving two-classed problems by determining the hyper-plane to separate two classes [111]. This is done by maximizing the margin from the hyper-plane to the bounds of two classes.

With the introducing of the kernel functions, SVM can be extended for non-linear problems by mapping the input data from low dimension to a higher dimension. Multi-class classification is also applicable, the multi-class SVM is built up by many two-class SVM networks either by using one-against-all or one-against-one method. The winning class is identified by the highest output function value or the maximum votes respectively [112].

In this section, we apply multi-class SVM classifiers to recognize the six basic facial expressions and neutral expression. SVMs are well suited for our system, since the high dimensionality of Gabor transformation does not affect training time of kernel classifiers. Given the training sets

$$T = \{(x_1, y_1), ..., (x_l, y_l)\} \in (X, Y)^l \tag{6.8}$$

where

$$x_i \in X = R, i = 1, 2, ..., l, y_i \in Y = \{+1, -1\} \tag{6.9}$$

it finds the solution of the following minimization problem during training

$$\min_{W,b,\xi} \frac{1}{2} W^T W + C_+ \sum_{y_i=+1} \xi_i + C_- \sum_{y_i=-1} \xi_i \tag{6.10}$$

subject to the ability of separating constraints

$$y_i (W^T \phi(x_i) + b) \geq 1 - \xi_i \quad (\xi_i \geq 0, i = 1, ..., l) \tag{6.11}$$

Suppose a set of training samples belong separate classes (x_1, y_1),…,(x_n, y_n) an optical hyperplane can be obtained by solving a constrained optimization problem. For a given kernel function, the SVM classify function can be stated as

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{l} y_i a_i K(x_i, x) + b) \tag{6.12}$$

where α_i (i=1,…,n) and b are the best optimization problem solution.

The most important part of SVMs is the selection of the kernel function. The kernel function may transform the data into a higher dimensional space to make it possible for performing the separation. We choose RBF as the kernel function:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0 \quad (6.13)$$

SVMs are described by the training samples and the kernel function.

When classifying seven facial expressions (anger, disgust, fear, happiness, sadness, surprise and neutral), a seven-class SVM classifier is performed. We apply one-against-one method, which is an effective way to solve multi-class problems by using multiple binary SVM classifiers. The one-against-one method constructs k (k-1)/2 classifiers where each one is constructed by using the training data from two classes chosen out of k classes. Each classifier is trained on a subset of the training set containing only training examples of the involved classes.

6.5 Experiment and Results

To evaluate the performance of the proposed method, we tested on the JAFFE database [115]. In our experiment, we use OpenCV as the software tool. The aim is to classify facial expressions into seven categories: happiness, sadness, surprise, angry, disgust, fear, and neutral respectively. The performance of the system is evaluated by the leave-one-subset-out cross-validation technique. We divided each category of facial expression into 3 subsets. Training and testing procedure was repeated 3 times for each category of facial expression. For each time, we picked out one subset out of the 3 subsets for the purpose of testing and the rest of the subsets used for training data. For each testing subset, we achieve a recognition rate. At the end, the overall recognition rate for each category of facial expression is the average value for all of the testing subsets. The recognition result for each category of facial expression is shown in Figure 6.7.

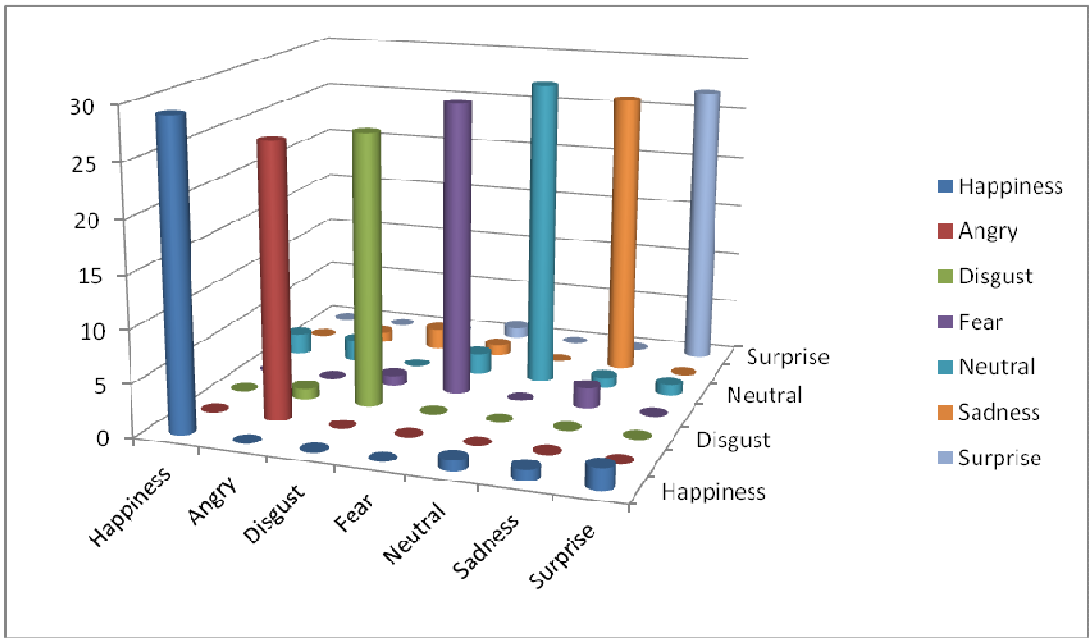


Figure 6.7 Recognition rate

Table 6.1 lists the classification results of all images for each category in the JAFFE database. For example, in the JAFFE database, there are 31 images for happiness facial expression. The classification results show that 29 out of the 31 images have been correctly recognized as happiness expression. However, 2 happiness images have been incorrectly recognized and categorized as neutral. For angry facial expression, 26 out of 30 images have been correctly recognized, 1 image has been incorrectly recognized as disgust, 2 images have been incorrectly recognized as neutral, and 1 image has been incorrectly recognized as sadness. As for the rest, the same logic follows.

Although the proposed method shows good recognition results by using the JAFFE database, there are some limitations of the method. First of all, the face image must be a frontal view. If a non-frontal-view image is given, the sum of the gray-scale value of the integral projection will change, resulted in the false detection of local minimum. Moreover, a non-frontal image will cause the hair connected with the eyebrows and eyes, therefore causing the false detection of local minimum. Second, since the purpose of the proposed system is for emotion detection for Chinese people, the thresholds we defined have only tested on Asian country images. The thresholds may need to be adjusted if tested on people from other ethnic (e.g. when the skin color is very dark).

Table 6.1 Recognition results

	Happiness	Angry	Disgust	Fear	Neutral	Sadness	Surprise
Happiness	29	0	0	0	2	0	0
Angry	0	26	1	0	2	1	0
Disgust	0	0	26	1	0	2	0
Fear	0	0	0	28	2	1	1
Neutral	1	0	0	0	29	0	0
Sadness	1	0	0	2	1	27	0
Surprise	2	0	0	0	1	0	27

6.6 Summary

This chapter recognized the six basic facial expression and neutral expression by using the integral projection techniques for image segmentation and Gabor filter for feature extraction. For emotion detection, only three facial expressions (happiness, angry, and sadness) are used as the input variables of the emotion model. In the future, if the psychology model is extended and more complex facial expression is needed in the model, the facial expression recognition method can be switched to facial muscle action (action unit) detection in order to recognized more complicated cases (e.g. non-frontal view facial expression recognition).

Chapter 7

Conclusion

7.1 Conclusion

The goal of this thesis is to investigate new solutions of emotion detection in HCI by integrating multi-channel information from the body language of the head in a natural and inexpensive way. This thesis enriches the current emotion recognition research in the following aspects:

1. When a facial expression is accompanied with other modalities from the body language of the head, entirely different emotions might be inferred. The implicit emotion is not always congruent with the explicit facial expression. However, this important factor has been ignored by current emotion detection systems. In this thesis, a two-stage approach is proposed. The first stage analyzes the explicit information from the modalities of facial expression, head movement, and eye gaze separately. In the second stage, all this information is fused to infer the implicit secondary emotional states. By integrating the channels from the body language of the head, the distinguished emotion may result in different quadrants in the emotional dimension space compared to the corresponding facial expression.

2. Head movement not only can provide information of the movement direction (head nod or head shake), but also can differ in speed and frequency (high frequency movement or low frequency movement). A very high frequency head movement may show much more arousal and active property than the low frequency head movement which differs on the emotion dimensional space. Although the frequency of head movement has a strong relationship with human emotion, current studies have not taken this important cue into consideration. This thesis examines emotional states not only by the direction of head movement, but also by the frequency of head movement. The head movement frequency is acquired by analyzing the tracking results of the coordinates from the detected nostril points. Emotional states under five categories of head movement including high frequency head nodding, low frequency head nodding, still, low frequency head shaking, and high frequency head shaking are studied.

3. In past research, eye gaze was considered in human behavioural recognition (e.g. the detection of human fatigue and attentive states). Psychologists found that gaze direction was associated with approach-oriented and avoidance-oriented emotions. These findings revealed that gaze direction influenced the processing of emotion displays. In this thesis, eye gaze direction is integrated with other head information to analyze emotional states. A geometrical relationship of human organs between nostrils and two pupils is developed to achieve this task. Four parameters are defined according to the changes in angles and the changes in the proportion of length of the four feature points to distinguish avert gaze from direct gaze. The sum of these parameters is considered as an evaluation parameter that can be analyzed to quantify gaze level.

4. New solutions are exploited for multimodality fusion by hybridizing the decision level fusion and the soft computing techniques to infer emotions. This could avoid the disadvantages of the decision level fusion technique, while retaining its advantages of adaptation and flexibility. However, there is no crisp boundary between high frequency head movement and low frequency head movement or direct gaze and averted gaze. How to quantify these modalities is another major concern of the thesis. A fuzzification strategy is proposed that can successfully quantify the extracted parameters of each modality into a fuzzified value between 0 and 1. These fuzzified values are the inputs for the fuzzy inference systems which

map the fuzzy values into emotional states.

7.2 Future Work

For future work, many possible improvements can be made to extend this work. First, more diversified emotional data samples can be collected from more people, and those samples can be used in the neural fuzzy training process so the training data set can fully represent the target group, providing more accurate results. The second improvement could be context-awareness for the emotion recognition system. The system could have a different solution of the fuzzy rules based on the context information of the user. For example, Chinese people and Indian people should have different fuzzy rules based on different cultural information. Other contexts, such as gender, age, etc. can also be taken into consideration. Since fuzzy systems are good in externality and adaptability, the soft computing approach again can be a good solution when adding the context information. Third, the facial expression recognition method can be switched to facial muscle action (action unit) detection because it is more flexible and extendable and easier to be fuzzified as well. Another possible improvement is to integrate more modalities into the inference system such as body posture, hand gesture, head tilt, etc. Adding more modalities could provide more information, thus providing possibilities to infer more complex emotional states. The human computer interaction experience can be enriched greatly and be much more interesting. Finally, more negative fuzzy rules can be added to the system to improve the system robustness.

The system developed in this work can be extended to many other research topics in the field of human computer intelligent interaction. We also can imagine that in the near future computers or robots will be intelligent enough to sense and response to the user's emotional states in a natural way. We hope this research will trigger more investigations to make computers friendlier, more intelligent, and more human-like.

Bibliography

[1] D. G. Myers, "Theories of Emotion," *Psychology: Seventh Edition*. NY: Worth Publishers, 2004.

[2] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang, "Human Computing and Machine Understanding of Human Behaviour: A Survey," in *Proc. Eighth ACM Int'l Conf. Multimodal Interfaces*, pp. 239-248, 2006.

[3] A. Kapoor, and R. W. Picard, "Multimodal affect recognition in learning environment," in *Proc. 13th Ann. ACM Int'l Conf. Multimedia*, pp. 677-682, 2005.

[4] P. Ekman, and W. V. Friesen, "Nonverbal behaviour in psychotherapy research," *Research in Psychotherapy*, vol. 3, pp. 179-216, 1968.

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

[6] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and Cognition*, Berlin: Springer-Verlag, vol. 17, no. 2, pp. 484–495. 2008.

[7] C. Darwin, *The Expression of the Emotions in Man and Animals*. NY: Oxford University Press, 1872.

[8] S. S. Tomkins, *Affect, Imagery, Consciousness: Vol. 1. The positive Affects*. NY: Springer, 1962.

[9] P. Ekman, “An argument for basic emotions,” *Cognition and Emotion*, vol. 6, pp. 169-200, 1992.

[10] P.Ekman, *Emotion in the human face*. UK: Cambridge University Press, 1982.

[11] A. Ortony, and T. J. Turner, “What’s basic about basic emotions?” *Psychological Review*, vol. 97, pp. 315–331, 1990.

[12] A. Wierzbicka, “Talking about emotions: Semantics, culture, and cognition,” *Cognition and Emotion*, vol. 6, pp. 3–4, 1992.

[13] S. Baron-Cohen, and T. H. E. Tead, *Mind reading: The interactive guide to emotion*. London: Jessica Kingsley Publishers, 2003.

[14] C. Yu, P. M. Aoki, and A. Woodruff, “Detecting user engagement in everyday conversations,” In *Proceedings of 8th International Conference on Spoken Language Processing*, (Jeju Island, Korea), pp. 1329-1332, 2004.

[15] C. Osgood, G. Suci, and P. Tannenbaum, *The measurement of meaning*. Chicago: University of Illinois Press, 1957.

[16] A. Mehrabian, and J. Russell, *An approach to environmental psychology*. MA: MIT Press, 1974.

[17] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[18] K. R. Scherer, A. Schorr, and T. Johnstone, (Eds.). *Appraisal processes in emotion: Theory, methods, research*, NY: Oxford University Press, 2001.

[19] K. R. Scherer, “Psychological models of emotion,” In *The neuropsychology of emotion* (J. Borod, eds.), NY: Oxford University Press, 2000.

[20] P. Ekman, “Expression and the nature of emotion,” *Approaches to Emotion*, pp. 319-344, 1984.

[21] P. Ekman, “Emotions inside out. 130 years after Darwin’s “The Expression of the Emotions in Man and Animal”,” *Ann. NY Acad. Sci.*, vol. 1000, pp. 1–6, 2003.

[22] D. Keltner, and P. Ekman, “Facial expression of emotion,” *Handbook of emotions* (M. Lewis and J. M. Haviland-Jones eds.), NY: Guilford Press, pp. 236-249, 2000.

[23] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: audio, visual, and spontaneous expressions,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, 2009.

[24] P. Ekman, “Facial expression and emotion,” *Am. Psychol.*, vol. 48, pp. 384–392, 1993.

[25] M. Pantic, and L. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-Computer Interaction," in *Proc. IEEE*, vol. 91, no. 9, pp. 1370-1390, 2003.

[26] L. Maat, and M. Pantic, "Gaze-X: Adaptive affective multimodal interface for single-user office scenarios," in *Proc. ACM Int'l Conf. Multimodal Interfaces*, pp. 171-178, 2006.

[27] J. F. Cohn, "Foundations of human computing: Facial expression and emotion," in *Proc. ACM Int'l Conf. Multimodal Interfaces*, pp. 233-238, 2006.

[28] P. Ekman, "The argument and evidence about universals in facial expressions of emotion," in *Psychological methods in criminal investigation and evidence*, (D. C. Raskin, eds.), pp. 297–332, 1989.

[29] P. Ekman, and W. V. Friesen, *Facial Action Coding System*. Palo Alto: Consulting Psychologist Press, 1978.

[30] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System*. Salt Lake City: A Human Face, 2002.

[31] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration: A statistical view," *IEEE Transactions on Multimedia*, vol. 1, no. 4, pp. 334–341, 1999.

[32] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully Automatic Facial Action Recognition in Spontaneous Behaviour," in *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pp. 223-230, 2006.

[33] J. F. Cohn, and K. L. Schmidt, "The Timing of Facial Motion in Posed and Spontaneous Smiles," in *Int'l J. Wavelets, Multiresolution and Information Processing*, vol. 2, pp. 1-12, 2004.

[34] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, and S. Kollias, "Emotion Recognition through Facial Expression Analysis Based on a Neurofuzzy Method," *Neural Networks*, vol. 18, pp. 423-435, 2005.

[35] S. Lucey, A. B. Ashraf, and J. F. Cohn, "Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face," *Face Recognition* (K. Delac, and M. Grgic, eds.), I-Tech Education and Publishing, pp. 275-286, 2007.

[36] N. Sebe, M. S. Lew, I. Cohen, Y. Sun, T. Gevers, and T.S. Huang, "Authentic Facial Expression Analysis," in *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2004.

[37] M. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, “Spontaneous versus Posed Facial Behaviour: Automatic Analysis of Brow Actions,” in *Proc. Eight Int’l Conf. Multimodal Interfaces*, pp. 162-170, 2006.

[38] M. F. Valstar, H. Gunes, and M. Pantic, “How to distinguish posed from spontaneous smiles using geometric features,” in *Proc. Ninth ACM Int’l Conf. Multimodal Interfaces*, pp. 38-45, 2007.

[39] Z. Zeng, Y. Fu, G. I. Roisman, Z. Wen, Y. Hu, and T. S. Huang, “Spontaneous Emotional Facial Expression Detection,” *J. Multimedia*, vol. 1, no. 5, pp. 1-8, 2006.

[40] G. C. Littlewort, M. S. Bartlett, and K. Lee, “Faces of Pain: Automated Measurement of Spontaneous Facial Expressions of Genuine and Posed Pain,” in *Proc. Ninth ACM Int’l Conf. Multimodal Interfaces*, pp. 15-21, 2007.

[41] J. F. Cohn, L. I. Read, Z. Ambadar, J. Xiao, and T. Moriyama, “Automatic analysis and recognition of brow actions and head motion in spontaneous facial behaviour,” in *Proc. IEEE Int’l Conf. Systems, Man, and Cybernetics*, vol. 1, pp. 610-616, 2004.

[42] A. Asthana, J. Saragih, M. Wagner, and R. Goecke, “Evaluating AAM Fitting Methods for Facial Expression Recognition,” in *Proc. 2009 Int’l Conf. Affective Computing and Intelligent Interaction*, 2009.

[43] T. Brick, M. Hunter, and J. Cohn, “Get the FACS Fast: Automated FACS Face Analysis Benefits from the Addition of Velocity,” in *Proc. 2009 Int’l Conf. Affective Computing and Intelligent Interaction*, 2009.

[44] M. E. Hoque, R. Kaliouby, R. W. and Picard, “When Human Coders (and Machines) Disagree on the Meaning of Facial Affect in Spontaneous Videos,” in *Proc. Ninth Int’l Conf. Intelligent Virtual Agents*, 2009.

[45] T. Johnstone, and K. R. Scherer, “Vocal Communication of Emotion,” *Handbook of Emotions*, pp. 220-235, 2000.

[46] C. M. Whissell, “The dictionary of affect in language,” in *Emotion: Theory, research and experience. The measurement of emotions.* vol. 4, pp. 113-131, (R. Plutchik, and H. Kellerman eds.), NY: Academic Press, 1989.

[47] N. Ambady, and R. Rosenthal, “Thin slices of expressive behaviour as predictors of interpersonal consequences: A meta–analysis,” *Psychological Bulletin*, vol. 11, no. 2, pp. 256–274, 1992.

[48] P. N. Juslin, and K. R. Scherer, "Vocal expression of affect," *The new handbook of methods in nonverbal behaviour research* (J. Harrigan, R. Rosenthal, and K. Scherer eds.), UK: Oxford University Press, pp. 65-135, 2005.

[49] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *Proceedings of the ISCA Workshop on Speech and Emotion,* (Belfast, Northern Ireland), pp. 19-24, Sept. 2000.

[50] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, and W. Fellenz, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32-80, 2001.

[51] J. A. Russell, and J. M. Fernández-Dols, *The psychology of facial expression.* NY: Cambridge University Press, 1997.

[52] M. Argyle, Bodily communication. London: Methuen, 1975.

[53] N. Hadjikhani, and B. De Gelder, "Seeing fearful body expressions activates the fusiform cortex and amygdale," *Current Biology*, vol. 13, pp. 2201-2205, 2003.

[54] R. A. Calvo, and S. D'Mello, "Affect detection: an interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect Computing*, vol. 1, pp. 18-37, 2010.

[55] N. Bernstein, *The Co-ordination and regulation of movements*. Oxford: Pergamon Press, 1967.

[56] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Nonverbal Behaviour*, vol.28, no. 2, pp. 117–139, 2004.

[57] J. Montepare, E. Koff, D. Zaitchik, and M. Albert, "The use of body movements and gesture as cues to emotions in younger and older adults," *J. Nonverbal Behaviour*, vol. 23, pp. 133-152, 1999.

[58] R. Walk, and K. Walters, "Perception of the smile and other emotions of the body and face at different distances," *Bull. Psychonomic Soc.*, vol. 26, pp.510, 1988.

[59] J. Van den Stock, R. Righart, and B. De Gelder, "Body expressions influence recognition of emotions in the face and voice," *Emotion (Washington, D.C.)*, vol. 7, no. 3, pp. 487-494, 2007.

[60] A. Kleinsmith, P. Ravindra De Silva, and N. Bianchi-Berthouze, "Grounding affective dimensions into posture features," in *Proceedings of the 1st International*

Conference on Affective Computing and Intelligent Interaction, (Beijing, China), pp. 263-270, Oct. 2005.

[61] S. Mota, and P. Picard, "Automated posture analysis for detecting learner's interest level," in *Proc. Computer Vision and Pattern Recognition Workshop*, vol. 5, p.49, 2003.

[62] S. D'Mello, and A. Graesser, "Automatic detection of learner's affect from gross body language," *Applied Artificial Intelligence*, vol. 23, pp. 123-150, 2009.

[63] R. W. Levenson, "Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity," *Social psychophysiology and emotion: Theory and clinical applications* (H. L. Wagner eds.), NY: John Wiley & Sons, pp. 17-42, 1988.

[64] A. Savran, K. Ciftci, G. Chanel, J. C. Mota, L. H. Viet, and B. Sankur, "Emotion detection in the loop from brain signals and facial images," in *Proceedings of eNTERFACE*, 2006.

[65] K. Takahashi, "Remarks on emotion recognition from multi-modal bio-potential signals," in *Proceedings of the IEEE International Conference on Industrial Technology*, (Hammamet, Tunisia), pp.1138-1143, 2004.

[66] J. Arroyo-Palacios, and D. M. Romano, "Towards a standardization in the use of physiological signals for affective recognition systems," in *Proceedings of Measuring Behaviour 2008,* (Maastricht, The Netherlands), pp. 121-124, 2008.

[67] R. A. Calvo, I. Brown, and S. Scheding, "Effect of Experimental Factors on the Recognition of Affective Mental States through Physiological Measures," in *Proc. 22nd Australasian Joint Conf. Artificial Intelligence*, 2009.

[68] O. AlZoubi, R. A. Calvo, and R. H. Stevens, "Classification of EEG for Affect Recognition: An Adaptive Approach," in *Proc. 22nd Australasian Joint Conf. Artificial Intelligence*, pp. 52-61, 2009.

[69] J. N. Bailenson, E. D. Pontikakis, I. B. Mauss, J. J. Gross, M. E. Jabon, C. A. C. Hutcherson, C. Nass, and O. John, "Real-Time Classification of Evoked Emotions Using Facial Feature Tracking and Physiological Responses," *Int'l J. Human-Computer Studies*, vol. 66, pp. 303-317, 2008.

[70] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Physiology-Based Affect Recognition for Computer-Assisted Intervention of Children with Autism Spectrum Disorder," *Int'l J. Human-Computer Studies*, vol. 66, pp. 662-677, 2008.

[71] A. Heraz, and C. Frasson, “Predicting the Three Major Dimensions of the Learner’s Emotions from Brainwaves,” *World Academy of Science, Eng. And Technology*, vol. 25, pp. 323-329, 2007.

[72] J. Wagner, N. J. Kim, and E. Andre, “From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification,” in *Proc. IEEE Int’l Conf. Multimedia and Expo*, pp. 940-943, 2005.

[73] K. Kim, S. Bang, and S. Kim, “Emotion Recognition System Using Short-Term Monitoring of Physiological Signals,” *Medical and Biological Eng. and Computing*, vol. 42, pp. 419-427, 2004.

[74] F. Nasoz, K. Alvarez, C. L. Lisetti, and N. Finkelstein, “Emotion Recognition from Physiological Signals Using Wireless Sensors for Presence Technologies,” *Cognition, Technology and Work*, vol. 6, pp. 4-14, 2004.

[75] A. Haag, S. Goronzy, P. Schaich, and J. Williams, “Emotion Recognition Using Bio-Sensors: First Steps towards an Automatic System,” *Affective Dialogue Systems*, pp. 36-48, 2004.

[76] C. E. Osgood, W. H. May, and M. S. Miron, *Cross-Cultural Universals of Affective Meaning*. Univ. of Illinois Press, 1975.

[77] C. Lutz, and G. White, “The Anthropology of Emotions,” *Ann. Rev. Anthropology*, vol. 15, pp. 405-436, 1986.

[78] J. A. Russell, “Core Affect and the Psychological Construction of Emotion,” *Psychological Rev.*, vol. 110, pp. 145-172, 2003.

[79] L. Barrett, “Are Emotions Natural Kinds?” *Perspectives on Psychological Science*, vol. 1, pp. 28-58, 2006.

[80] M. A. Cohn, M. R. Mehl, and J. W. Pennebaker, “Linguistic Markers of Psychological Change Surrounding September 11, 2001,” *Psychological Science*, vol. 15, pp. 687-693, Oct. 2004.

[81] J. Pennebaker, M. Mehl, and K. Niederhoffer, “Psychological Aspects of Natural Language Use: Our Words, Our Selves,” *Ann. Rev. Psychology*, vol. 54, pp. 547-577, 2003.

[82] J. Kahn, R. Tobin, A. Massey, and J. Anderson, “Measuring Emotional Expression with the Linguistic Inquiry and Word Count,” *Am. J. Psychology*, vol. 120, pp. 263-286, 2007.

[83] C. G. Shields, R. M. Epstein, P. Franks, K. Fiscella, P. Duberstein, S. H. McDaniel, and S. Meldrum, "Emotion Language in Primary Care Encounters: Reliability and Validity of an Emotion Word Count Coding System," *Patient Education and Counseling*, vol. 57, pp. 232-238, 2005.

[84] Y. Bestgen, "Can Emotional Valence in Stories Be Determined from Words," *Cognition and Emotion*, vol. 8, pp. 21-36, Jan. 1994.

[85] J. Hancock, C. Landrigan, and C. Silver, "Expressing Emotion in Text-Based Communication," in *Proc. SIGCHI*, 2007.

[86] J. Pennebaker, M. Francis, and R. Booth, *Linguistic Inquiry and Word Count (LIWC): A Computerized Text Analysis Program*. Erlbaum Publishers, 2001.

[87] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Introduction to Wordnet: An On-Line Lexical Database," J. Lexicography, vol.3, pp. 235-244, 1990.

[88] S. D'Mello, N. Dowell, and A. Graesser, "Cohesion Relationships in Tutorial Dialogue as Predictors of Affective States," in *Proceedings of the 2009 conference on Artificial Intelligence in Education*, 2009.

[89] T. Danisman, and A. Alpkocak, "Feeler: Emotion Classification of Text Using Vector Space Model," in *Proc. AISB 2008 Convention, Comm., Interaction and Social Intelligence*, 2008.

[90] C. Strapparava, and R. Mihalcea, "Learning to Identify Emotions in Text," in *Proc. 2008 ACM Symp Applied Computing*, pp. 1556-1560, 2008.

[91] W. H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, "Which Side Are You On? Identifying Perspectives at the Document and Sentence Levels," in *Proc. 10th Conf. Natural Language Learning*, pp. 109-116, 2006.

[92] C. O. Alm, D. Roth, and R. Sproat, "Emotions from Text: Machine Learning for Text-Based Emotion Prediction," in *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347-354, 2005.

[93] Z. Zeng, Y. Hu, G. Roisman, Z. Wen, Y. Fu, and T. S. Huang "Audio-visual emotion recognition in adult attachment interview," in *Proc. Int'l Conf. Multimodal Interfaces*, (J. Y. F. K. H. Quek, D. W. Massaro, A. A. Alwan, and T. J. Hazen, eds.), pp. 139-145, 2006.

[94] A. Corradini, M. Mehta, N. O. Bernsen, and J. C. Martin, "Multimodal input fusion in human computer interaction on the example of the ongoing nice project," in *Proceedings of the NATO*, (Tsakhkadzor, Armenia), pp. 223-234, 2003.

[95] K. Scherer, and H. Ellgring, "Multimodal expression of emotion: affect programs or componential appraisal pattern?" *Emotion*, vol.7, pp 158-171, 2007.

[96] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," *Affect and Emotion in Human-Computer Interaction*, pp. 92-103, 2008.

[97] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, "Emotion sensors go to school," in *Proc. 14th Conf. Artificial Intelligence in Education*, pp. 17-24, 2009.

[98] A. Kapoor, B. Burleson, and R. W. Picard, "Automatic Prediction of Frustration," *Int'l J. Human-Computer Studies*, vol. 65, pp. 724-736, 2007.

[99] S. D'Mello, and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 10, pp 147-187, 2010.

[100] G. Caridakis, K. Karpouzis, and S. Kollias, "User and context adaptive neural networks for emotion recognition," *Neurocomputing*, vol. 71, pp. 13-15, 2553-2562, 2008.

[101] J. Kim, "Bimodal emotion recognition using speech and physiological changes," *Robust speech recognition and understanding* (M. Grimm, and K. Kroschel eds.), Vienna, Austria: I-Tech Education and Publishing, pp. 265-280, 2007.

[102] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, and L. Malatesta, "Modelling naturalistic affective states via facial, vocal and bodily expressions recognition," *Artifical Intelligence for Human Computing: ICMI 2006 and IJCAI 2007 International Workshops*, (J. G. Carbonell, and J. Siekmann eds.), Banff, Canada, pp. 92-116, 2007.

[103] D. Kulic, and E. A. Croft, "Affective state estimation for human–robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.

[104] R. E. Kaliouby, and P. Robinson, "Generalization of a Vision-Based Computational Model of Mind-Reading," in *Proc. First Int'l Conf. Affective Computing and Intelligent Interaction*, pp. 582-589, 2005.

[105] P. Ekman, and W. V. Friesen, *Unmasking the face*. NJ: Prentice-Hall, 1975.

[106] P. Viola, and M. Jones, ''Robust real-time face detection,'' *International Journal of Computer Vision*, vol.57, no. 2, pp. 137-154, 2004.

[107] P. Viola, and M. Jones, ‘‘Robust real-time object detection,’’ *Cambridge Research Laboratory Technical Report Series*, pp. 1-24, 2001.

[108] L. Shen, and L. Bai, “A review on Gabor wavelets for face recognition”, *Pattern Anal Applic*, vol. 9, pp. 273-292, 2006.

[109] C. Burges, “A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[110] N. Cristianini, and J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods.* Cambridge University Press, 2006.

[111] V. Vapnik, *The nature of statistical learning theory*. Springer Verlag, 2000.

[112] G. Fung, and O. Mangasarian, “Multicategory proximal support vector machine classifiers,” *Machine Learning*, vol. 59, no. 1, pp. 77-97, 2005.

[113] J. Shawe-Taylor, and N, Cristianini, *Kernel methods for pattern analysis.* Cambridge University Press, 2004.

[114] M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28, 1998.

[115] M. J. Lyons, J. Budynek, and S. Akamatsu, “Automatic classification of single facial images,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357-1362, 1999.

[116] D. Vukadinovic, and M. Pantic, “Fully automatic facial feature point detection using Gabor feature based boosted classifiers,” *IEEE Int’l Conf. Systems, Man, and Cybernetics*, vol. 2, pp 1692 – 1698, 2005.

[117] C. Harris, and M. Stephens, “A combined corner and edge detector,” in *Proceedings of the 4th Alvey Vision Conference*, vol. 15, pp. 146-151, 1988.

[118] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting,” *Annals of Statistics*, vol. 28, no. 2, pp. 337-374, 2000.

[119] J. Jones, and L. Palmer, “An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex,” *J. Neurophysiology*, vol. 58, no. 6, pp. 1233-1258, 1978.

[120] J. Shi, and C. Tomasi, “Good features to track,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pp. 593-600, 1994.

[121] E. O. Brigham, *The fast Fourier transform*. NY: Prentice-Hall, 2002.

[122] A. Freitas-Magalhães, *The Psychology of Emotions: The Allure of Human Face.* Oporto: University Fernando Pessoa Press, 2007.

[123] J. D. Rothwell, *In the Company of Others: An Introduction to Communication.* US: McGraw-Hill, 2004.

[124] S. S. Al-amri, N. V. Kalyankar, and S. D. Khamitkar, "Linear and non-linear contrast enhancement image," *International Journal of Computer Science and Network Security*, vol. 10, no. 2, pp. 139-143, 2010.

[125] T. Kanade, J. Cohn, and Y. Tian. "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46-53, 2000.

[127] B. Reginald, Jr. Adams, and E. K. Robert, "Effects of direct and averted gaze on the perception of facially communicated emotion," *Emotion*, vol. 5, No. 1, pp, 3-11, 2005.

[128] T. Ganel, "Revisiting the relationship between the processing of gaze direction and the processing of facial expression," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 1, pp. 48-57, 2011.

[129] B. Reginald, Jr. Adams, G. Robert, and Jr. Franklin, "Influence of emotional expression on the processing of gaze direction," *Motivation and Emotion*, vol. 33, no. 2, pp 106-112, 2009.

[130] M. Pantic, and M. S. Bartlett, "Machine analysis of facial expressions," *Face recognition.* Vienna, Austria: I-Tech Education and Publishing. pp. 377-416, 2007.

[131] H. Gunes, and M. Piccardi, "From monomodal to multi-modal: affect recognition using visual modalities," *Ambient intelligence techniques and applications*, Berlin: Springer-Verlag, pp. 154-182, 2009.

[132] R. E. Kaliouby, and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Proc. Int'l Conf. Computer Vision and Pattern Recognition*, vol. 3, pp. 154, 2004.

[126] J. Borg, *Body Language: 7 Easy Lessons to Master the Silent Language*. FT Press, 2010.

[133] S. Baron-Cohen, O. Golan, S. Wheelwright, and J. Hill. *Mind Reading: The Interactive Guide to Emotions.* London: Jessica Kingsley Publishers, 2004.

[134] Q. Ji, P. Lan, and C. Looney “A probabilistic framework for modeling and real-time monitoring human fatigue,” *IEEE Systems, Man, and Cybernetics Part A*, vol. 36, no. 5, pp. 862-875, 2006.

[135] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, “Estimation of behavioural user state based on eye gaze and head pose – application in an e-learning environment,” *Journal in Multimedia Tools and Applications*, vol. 41, 2008.

[136] A. Gegov, “Complexity management in fuzzy systems,” *Studies in Fuzziness and Soft Computing*, vol. 211, 2007.

[137] L. Zadeh, “Outline of a new approach to the analysis of complex systems and decision processes,” *IEEE Trans. Syst., Man, Cybern.*, vol. 3, no. 1, pp. 28-44, 1973.

[138] J. S. R. Jang, “ANFIS: adaptive-network-based fuzzy inference systems,” *IEEE Transaction Systems, Man and Cybernetics*, vol. 23, pp. 665-685, 1993.

[139] J. S. Taur, and C. W. Tao, “A new neuro-fuzzy classifier with application to on-line face detection and recognition,” *Journal of VLSI Signal Processing*, vol. 26, no. 3, pp. 397-409, 2000.

[140] M. Khezri, M. Jahed, N. Sadati, “Neuro-fuzzy surface EMG pattern recognition for multifunctional hand prosthesis control,” *IEEE International Symposium on Industrial Electronics*, pp. 269-274, 2007.

[141] M. Engin, “ECG beat classification using neuro-fuzzy network,” *Pattern Recognition Letters*, vol. 25, pp. 1715-1722, 2004.

[142] M. Khezir, M. Jahed, “Real-time intelligent pattern recognition algorithm for surface EMG signals,” *BioMedical Engineering Online*, vol. 6, 2007.

[143] G. Feng, “A survey on analysis and desgin on model-based fuzzy control system,” in *IEEE Transaction on Fuzzy Systems*, vol. 14, no. 5, 2006.

[144] A. Chakraborty, and A. Konar, “Emotion recognition from facial expressions and its control using fuzzy logic,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 39, no. 4, 2009.

[145] R. Contreras, O. Starostenko, V. Alarcon-Aquino, and L. Flores-Pulido, “Facial feature model for emotion recognition using fuzzy reasoning,” *Advances in Pattern Recognition*, vol. 6256, pp. 11-21, 2010.

[146] N. Esau, E. Wetzel, L. Kleinjohann, and B. Kleinjohann, “Real-time facial expression recognition using a fuzzy emotion model,” in *IEEE Int. Conf. Fuzzy Systems*, pp. 351-356, 2007.

[147] R. L. Mandryk, and M. S. Atkins, “A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies,” *International Journal of Human-Computer Studies*, vol. 65, pp. 329-347, 2007.

[148] S. Chatterjee, and S. Hao, “A novel neuro fuzzy approach to human emotion determination,” in *International Conference on Digital Image Computing: Techniques and Applications*, pp. 283-287, 2010.

[149] S. Ioannou, G. Caridakis, K. Karpouzis, and S. Kollias, “Robust feature detection for facial expression recognition,” *Journal on Image and Video Processing*, vol. 2007, 2007.

[150] C. D. Katsis, N. Katertsidis, G. Ganiatsas, and D. I. Fotiadis, “Toward emotion recognition in car-racing drivers: a biosignal processing approach,” *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* vol. 38, no. 3, pp. 502- 512, 2008.

[151] C. Lee, and S. Narayanan, “Emotion recognition using a data-driven fuzzy inference system,” *Eurospeech*, pp. 157–160, 2003.

[152] S. Giripunje, and N. Bawane, “ANFIS based emotions recognition in speech,” *Springer-Verlag*, vol. 4692, pp. 77-84, 2007.

[153] J. Schalk, S. T. Hawk, A. H. Fishcher, and B. Doosje, “Moving faces, looking places: validation of the Amsterdam dynamic facial expression set (ADFES),” *Emotion*, vol. 11, no. 4, pp. 907-920, 2011.

[154] A. A. Marsh, H. A. Elfenbein, and N. Ambady, “Nonverbal “accents”: cultural differences in facial expressions of emotion,” *Psychological Science*, vol. 14, no. 4, pp. 373-376, 2003.

[155] D. Matsumoto, “Cultural influences on the perception of emotion,” *Journal of Cross-Cultural Psychology*, vol. 20, no. 1, pp. 92-105, 1989.

[156] D. Matsumoto, *The handbook of culture and psychology*. Oxford University Press, 2001.

[157] G. Geri-Ann, *Caring for Patients from Different Cultures*. University of Pennsylvania Press, 2004.

[158] J. Lukasiewicz, “On 3-valued logic,” *Ruch Filozoficzny*, (Polish), vol. 5, pp. 169-171, 1920.

[159] T. Takagi, and M. Sugeno, “Fuzzy identification of systems and its applications to modeling and control,” *IEEE Trans Syst Man Cybern*, vol. 15, no. 1, pp. 116-132, 1985.

[160] E. H. Mamdani, and S. Assilian, “An experiment in linguistic synthesis with a fuzzy logic controller,” *Int. J. Man Mach. Stud.*, vol. 7, pp. 1–13, 1975.

[161] S. O. Ba, and J. M. Odobez, “Recognizing visual focus of attention from head pose in natural meetings,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 39, 2009.

[162] E. C. Vural, E. Mujdat, A. Ercil, G. Littlewort, M. Bartlett, M. Marian and J. Movellan, “Automated drowsiness detection for improved driving safety,” in *International Conference on Automotive Technologies*, (Istanbul), 2008.

[163] O. Villon, and C. Lisetti, “A User-Modeling Approach to Build User’s Psycho-Physiological Maps of Emotions Using Bio- Sensors,” in *Proc. IEEE 15th Int’l Symp.*, pp. 269-276, 2006.

[164] R. Sharma, V. I. Pavlovic, and T. S. Huang, “Toward Multimodal Human-Computer Interface,” in *Proc. IEEE*, vol. 86, no. 5, pp. 853- 869, 1998.

[165] P. Ekman, and W.V. Friesen, *Unmasking the face*, Prentice-Hall, New Jersey, USA，1975.

[166] E.M. Chutorian, and M.M. Trivedi, “Head Pose Estimation in Computer Vision: A Survey,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4. 2009.

[167] C. Merten, and C. Conati, “Eye Tracking to Model and Adapt to User Meta-cognition in Intelligent Learning Environments,” in *Proceedings of the 11th international Conference on intelligent User interfaces*, pp. 39 – 46. 2006.

[168] K. Nguyen, C. Wagner, D. Koons, and M. Flickner, “Differences in the Infrared Bright Pupil Response of Human Eye,” in *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, pp. 133–138, 2002.

[169] B. Noris, K. Benmachiche, and A. Billard, “Calibration-Free Eye Gaze Direction Detection with Gaussian Processes,” in *Proceedings of the International Conference on Computer Vision Theory and Applications*, 2008.

[170] T. Ohno, N. Mukawa, and A. Yoshikawa, “FreeGaze: A Gaze Tracking System for Everyday Gaze Interaction,” in *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, pp. 125–132, 2002.

[171] D. Li, J. Babcock, and D.J. Parkhurst, “OpenEyes: A Low-cost Head-mounted Eye-tracking Solution,” in *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, pp. 95–100, 2006.

[172] A. Bulling, D. Roggen, and G. Tröster, “EyeMote – Towards Context-Aware Gaming Using Eye Movements Recorded From Wearable Electrooculography,” in *Proc. of the 2nd International Conference on Fun and Games*, Springer, pp. 33–45, 2008.

[173] A. De Luca, R. Weiss, and H. Drewes, “Evalution of Eye-Gaze Interaction methods for Security Enhanced PIN-Entry,” in *Proceedings of the 19th Australasian Conference on Computer-Human interaction*, vol. 51, pp. 199–202, 2007.

[174] M. Ashmore, A.T. Duchowski, and G. Shoemaker, “Efficient Eye Pointing with a Fisheye Lens,” in *Proceedings of Graphics interface*, vol. 112, pp. 203–210, 2005.

[175] D. Beymer, and D.M. Russell, “WebGazeAnalyzer: A System for Capturing and Analyzing Web Reading Behaviour Using Eye Gaze,” in *Extended Abstracts on Human Factors in Computing Systems*, pp. 1913–1916, 2005.

[176] G.T. Böning, K. Bartl, T. Dera, S. Bardins, E. Schneider, and T. Brandt, “Mobile Eye Tracking as a Basis for Real-Time Control of Gaze-Driven Head-Mounted Video Camera,” in *Proceedings of the Eye Tracking Research & Applications Symposium*, 2006.

[177] H. Drewes, A. De Luca, and A. Scgnudt, “Eye-Gaze Interaction for Mobile Phone,” in *Proceedings of the 4th international Conference on Mobile Technology Mobility '07*, pp. 364–371, 2007.