

PROTECTED MULTIMODAL EMOTION RECOGNITION

by

Kevin Tang

Bachelor of Electrical Engineering, Ryerson University, Toronto, Canada 2011

A Thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Applied Science

in the Program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2015

©Kevin Tang 2015

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research

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

I understand that my thesis may be made electronically available to the public.

Protected Multimodal Emotion Recognition

Master of Applied Science 2015

Kevin Tang

Electrical and Computer Engineering

Ryerson University

Abstract

In this thesis, we propose Protected Multimodal Emotion recognition (PMM-ER), an emotion recognition approach that includes security features against the growing rate of cyber-attacks on various databases, including emotion databases. The analysis on the frequently used encryption algorithms has led to the modified encryption algorithm proposed in this work. The system is able to recognize 7 different emotions, i.e. happiness, sadness, surprise, fear, disgust, and anger, as well as a neutral emotion state, based on 2D video frames, 3D vertices, and audio wave information. Several well-known features are employed, including the HSV colour feature, iterative closest point (ICP) and Mel-frequency cepstral coefficients (MFCCs). We also propose a novel approach to feature fusion including both decision- and feature-level fusion, and some well-known classification and feature extraction algorithms such as principle component analysis (PCA), linear discernment analysis (LDA) and canonical correlation analysis (CCA) are compared in this study.

Acknowledgements

I would like to thank my supervisor, Dr. Ling Guan, for his help and support. Without him, I would never have been able to enter this field of enjoyable research. I also would like to thank my co-supervisor, Dr. Truman Yang, for his help and support. Also, I would like to thank my mentor, Dr. Yun Tie, Dr. Rui Zhang and Miss. Guo Xin, for their guidance and support. I am glad to complete this work of multimodal recognition system because of you and everyone who has supported me. Include my lovely family, especially my mother Mei Zu Chen, I love this research, and I wish to share my gratitude with you in this thesis.

Contents

Author's Declaration	ii
Abstract	iii
Acknowledgements	iv
Contents	v
List of Tables	vii
List of Figures	viii
1 Introduction.....	1
1.1 Background	1
1.2 Literary review	2
1.3 Challenge	3
1.4 Outline.....	6
2 System Overview	8
2.1 System overview	9
2.2 Contributions.....	12
3 A Novel Multimodal Emotion Recognition Framework	14
3.1 Data Collection	15
3.1.1 Device Selection	15
3.1.2 The Recording Procedure	19
3.2 Pre-processing.....	23
3.3 Feature extraction.....	35
3.3.1 2D image based feature extraction.....	37
3.3.2 3D vertices based feature extraction	39
3.3.3 Audio based feature extraction	42
3.3.4 Dimensionality reduction.....	44
3.4 Proposed fusion method.....	46
3.5 Classification.....	49

4 Database and security	51
4.1 Database	51
4.2 Security - design	52
4.3 Overview of database	55
5 Experiment	57
5.1 Experiment - Security	57
5.2 Recognition accuracy	59
5.2.1 2D based recognition	62
5.2.2 Audio based recognition	64
5.2.3 3D Feature based recognition	66
5.2.4 All three features based recognition (PMM-ER)	69
5.3 Summary	71
6 Conclusion	72
6.1 Summary	72
6.2 Future research	73
Reference	76
List of publication	83

List of Tables

Table 2.1 Characteristics of Each Data Type	12
Table 3.1 The schematic description of recording procedure	20
Table 3.2 Languages analysis on audio	34
Table 3.3 Audio Features	43
Table 5.1. Encryption methods comparison on the data.	58
Table 5.2 Recognition accuracy vs. trained	61
Table 5.3 Recognition accuracy (2D)	62
Table 5.4 Recognition accuracy (ratio = 8).....	63
Table 5.5 Audio feature in time, log and MFCC spectrum.....	64
Table 5.6 Recognition accuracy (audio)	67
Table 5.7 3D emotion modal example	68
Table 5.8 Recognition accuracy (3D)	69
Table 5.9 Recognition accuracy (2D + 3D + Audio)	69
Table 5.10 Recognition accuracy for other research.....	70

List of Figures

Figure 2.1 Project outline.....	9
Figure 3.1 Kinect.....	16
Figure 3.2 Kinect IR camera data	17
Figure 3.3 KINECT horizontal view angle range	18
Figure 3.4 KINECT vertical view angle range	18
Figure 3.5 Collection procedure.....	21
Figure 3.6 (a) Device set-up position. (b) chair position	22
Figure 3.7 2D image data in database	24
Figure 3.8 2D pre-processing using filters.....	25
Figure 3.9 Emotion of disgust.....	25
Figure 3.10 The installation of an extra lens.....	26
Figure 3.11 (a) Result without lens and (b) with lens	28
Figure 3.12 Example of the Kinect face tracking point position	28
Figure 3.13 (a) before 3D pre-process (b)after 3D pre-process	30
Figure 3.14 Pre-process for emotion audio.....	31
Figure 3.15 Language relationship.....	33
Figure 3.16 Compare between Korean (left) with Persian (right)	35
Figure 3.17 2D Feature extraction and recognition process	36
Figure 3.18 HSV colour modal	38
Figure 3.19 Global and local ICP.....	41
Figure 3.20 Classification process	46
Figure 3.21 Decision level fusion.	47
Figure 3.22 Feature level fusion.	48
Figure 4.1. Cryptography.....	53
Figure 4.2 Shifting to create disorder.....	54

Figure 5.1. Result of encryption time.	58
Figure 5.2 2D emotions frames.....	60
Figure 5.3 Recognition accuracy vs. trained	62
Figure 5.4 PCA + LDA recognition accuracy.....	63
Figure 5.5 Comparison of recognition results.....	71
Figure 6.1. New KINECT model (KINECT for Xbox360 and XBoxONE).....	75

Chapter 1

Introduction

1.1 Background

Since the dawn of the so-called computer age, digital information has become the standard technology for data storage, and the use of computers has become an integral aspect of our daily lives. In this environment, human-computer interaction (HCI) underpins all data processing. The implementation of accurate, “user-friendly”, and secure HCI processes is the key for human mastery of digital information. There are many different levels of HCI, ranging from highway electronic toll collection (ETC) systems, through vending machines in schools, to iris recognition machines in airports [7]. In recent years, there has been a greater number of camera-based computer-vision HCI applications that essentially mimic the human brain, through which the computer can “see” or process information, including iris, fingerprint, gesture, and emotion recognition applications. The key factor of any recognition system is the recognition accuracy, and a system with low recognition accuracy has little chance for success.

From a philosophical point of view, emotions are expression of our thoughts. In other words, emotions are the outcome of our rational thinking. Different people have different thoughts, and thus the resulting emotions are equally different. An example of such ambivalent emotional expression is Leonardo da Vinci’s iconic *La Gioconda*, better known as the *Mona Lisa*, a portrait based on a real woman’s facial expression that evokes a common question: is she happy or sad? While the original artwork does not clearly indicate the model’s happiness, the literature shows

that most people perceive the titular subject as being happy, though many would argue that she is a bit melancholic, if not sad. Such a debate illustrates how we, or more specifically how our brains, classify facial expressions. Recent studies have shown that some emotions are controlled by separate brain hemispheres. For example, the right hemisphere decides how emotions can be expressed, but the left hemisphere has the duty of decision making. That means that although individuals may be happy, they may not act as happy as they think. In short, with so many different variables, and so many different emotion types, it is challenging to determine the starting point for the study of human emotion.

American psychologist Paul Eckman has suggested that humans have six basic emotions that are universal throughout human cultures: fear, disgust, anger, surprise, happiness, and sadness [5]. To place it into a visual context, we have the primary colours of red, blue, and yellow; combining yellow and red will result in orange. The basic emotions behave in a similar way. Based on this concept, the main focus of this research is the classification of human facial expression into these six basic emotions, including happy, sadness, angry, fear, disgust and surprise plus neutral emotion, total of seven emotion types used in our work.

1.2 Literature Review

Emotion recognition is a process performed by human or computer system, this process including extracting emotion feature from the object (a person) and classifying the emotion type (e.g. six basic emotions) those emotion features belong to [47]. On regular daily basis, we recognize other people's emotion based on their facial expression, speech, body language and many other physically and mental expression from a person. Generally speaking, most of the emotion recognition involve visual and audio based data [34]. On the visual side, a 2D facial image can be

used to describe the emotion of a person. One of a well-known analysis method is called facial action coding system (FACS) [37]. The original FACS been proposed by P. Ekman in 1978, which defines 44 facial expression action units (AUs). AU has been used to describe how emotion is expressed. For example, based on the general society according to the Introductory Psychology in Yale University [38], a very simple smiley face would represent happy to our brain. This would be a method of translate from visual to content that smiley face can be translated as happy without using any actual language.

Reviewing the current visual based emotion recognition research shows that AU is helpful in the emotion recognition, in which the number of AU classes is generally between 12 and 22 [38]. But most of them still use the six basic emotions as the classes [39]. On the algorithm side, with the use of classifier algorithms such as the deep belief networks (DBN), Tong and others has achieved the recognition accuracy of 88.3% in [40]. An improved DBN with result of recognition accuracy of 94.05% by Li and others [41] employs the gentles support vector machine (GSVM) with a high recognition accuracy as 95.3% [42]. The algorithm of using template matching achieved 88.33% recognition accuracy [43]. On the audio part, the recognition accuracy varies from 62.8%+6.69% [44] to as high as 91.55% [45] depending on the number of classes, the quality of the database, as well as the recognition methods. Most of the audio emotion recognition approaches use hidden Markov models (HMMs). The classes are similar again, i.e. the six basic emotions are still the common classes.

1.3 Challenges

Research on multimodal emotion recognition was inspired by the desire to create a highly accurate emotion recognition system that is able to reflect a real emotion using computer technology. As

mentioned earlier, many factors affect the accuracy of emotion recognition. For example, a low resolution camera may produce inferior results compared to a high resolution camera. The quality of voice recording in a noisy environment may be lower than that in a quiet place. An unsecured database can result in possible hacking to the database itself. Finally, an inefficient algorithm may take longer to execute and/or lead to lower recognition accuracy. Beyond that, the research also focuses on finding real emotion, which is the most accurate reflection of the emotion that a person may be experiencing and expressing. For instance, a person crying with a mumbling voice may smile at the same time. Determining if a person is experiencing greater happiness or sadness requires the consideration of multiple factors.

According to Mehrabian [46], the communication of human emotion comprises the information which is 7% verbal, i.e. the speech, 38% vocal, i.e. the tone of the word(s), and 55% visual, which includes facial expressions and body language. Although facial expression represents the most significant factor in human emotional expression, we might not be able to accurately and fully identify individuals' real emotion without adequately considering emotional expression's verbal and vocal components. In pursuit of such accurate identification, we introduce a multimodal emotion database that is designed to collect individuals' real emotion.

In this study, we considered 2D facial images, 3D feature points, also referred to as fiducial points or vertices points, and audio data. By designing a new emotion collecting procedure within a consistent environment and using a selected device, we have built a multimodal emotion database. On the fusion part, we have tested the decision- and feature-level fusion and also propose our new fusion algorithms, particularly to improve the emotion recognition accuracy.

The lighting condition used in the recording environment, the angle of the user's face, and the background noise during audio recording are all factors that potentially affect the performance of emotion recognition. The four major challenges of emotion recognition are:

1. Device selection
2. Environment
3. Security
4. Algorithm selection and design

Device selection. The first challenge is choosing a suitable device. The goal is to find a device that can record red green blue colour modal (RGB) video, 3D vertices, and audio simultaneously. Although it is possible to build such a system using separate devices, such as the combination of an RGB video recorder and a microphone, such dual devices may lead to the difficulty in coordinating the real-time processes; in other words, the raw video may not be synchronized exactly with the audio recording. An even more significant problem may occur when coordinating RGB video and 3D vertices; the emotion 2D / 3D / audio data should be recorded from a fixed point with the same distance and angle to the person being recorded to provide an identical frame size for both the 2D RGB video and the 3D feature point data. Therefore, in order to provide such an environment, we use a single device that is able to simultaneously record 2D RGB video and 3D feature points instead of combining two or more devices. Both ASUS XTON and Microsoft KINECT satisfy such criteria, and therefore are most suitable for our project as will be discussed in a subsequent chapter.

Environment. Feature selection is done after the data are collected and pre-processed. This step is extremely important since the recognition accuracy depends on the quality of the data. The quality of the data directly depends on the quality of recording environment.

Security. To overcome the problem of potential attacks on the database, such as SQL injection, we have designed a suitable protection scheme using encryption. There are three different file types for the database, i.e. the text file (.txt), the audio video interleave file (.avi), and the windows media audio file (.wma). In this study, we analyze each file in terms of its file size, encryption time, decryption time, transfer rate, and error rate in order to determine which encryption is best suited for our database. We also test the three most frequently used encryption algorithms: Data Encryption Standard (DES), Triple Data Encryption Algorithm (3DES), and Advanced Encryption Standard (AES). Based on the experiment results, AES is the most reliable, trust-worthy and fastest among the three encryption algorithms.

Algorithm. Fusion is the last step of the recognition process, which combines multiple information streams. In this study, we have researched and modified the most suitable fusion algorithm for multimodal emotion recognition.

1.4 Outline

The rest of the thesis is organized as follows.

Chapter 2 presents an overview of the proposed multimodal emotion recognition system, including the architecture of the system and the main contributions.

Chapter 3 discusses data collection, pre-processing, feature extraction and multi-modal emotion recognition. We first provide details on the device selection, raw data processing, data collection methodology, pre-processing on 2D, 3D, and audio data, 2D image feature extraction, 3D vertices feature extraction, and audio feature extraction. Then, we discuss the methodology for the emotion recognition, including fusion at both decision and feature levels, as well as our proposed fusion method for the recognition purpose.

Chapter 4 outlines the system's security features and its design. In addition, we compare Data Encryption Standard (DES), Triple Data Encryption Algorithm (3DES), and Advanced Encryption Standard (AES) encrypting / decrypting algorithms.

Chapter 5 presents the results of the experiment, including the comparisons with respect to the database, security, fusion, and recognition.

Chapter 6 draws conclusions and point out some directions for future research.

Chapter 2

System Overview

The proposed multimodal emotion recognition procedure builds on three different data types: 2D video image frame, 3D feature points, and audio data. Each type can be used independently to achieve around 55%+ recognition accuracy for either human face identification or human visual emotion recognition [32].

Most recognition systems use only a single type of data. For example, a fingerprint recognition system only needs an image of fingerprint. On the other hand, multimodal recognition systems involve combination of different types of input data. For dimensionality reduction and fusion, the method includes principal component analysis (PCA), linear discriminant analysis (LDA) and canonical correlation analysis (CCA). Although PCA is the most common one to use, in case of multimodal data, CCA works better than PCA. On the other hand, the literature reports on state-of-the-art classification methods including but not limited to the nearest neighbour algorithm, support vector machines (SVMs), neural networks and hidden Markov models (HMMs). Each classification method is suitable for particular data types depend on the different cases. For example, HMMs work better for audio data in terms of recognition accuracy when compared to nearest neighbour. The selection and improvement of algorithms not only increase accuracy but is also meant to suit the size of the database as well as to improve efficiency and cost. Most methods involving fusion of all three types of data focus on the decision-level fusion. In order to address the drawbacks of existing fusion methods, improve recognition results and prevent possible

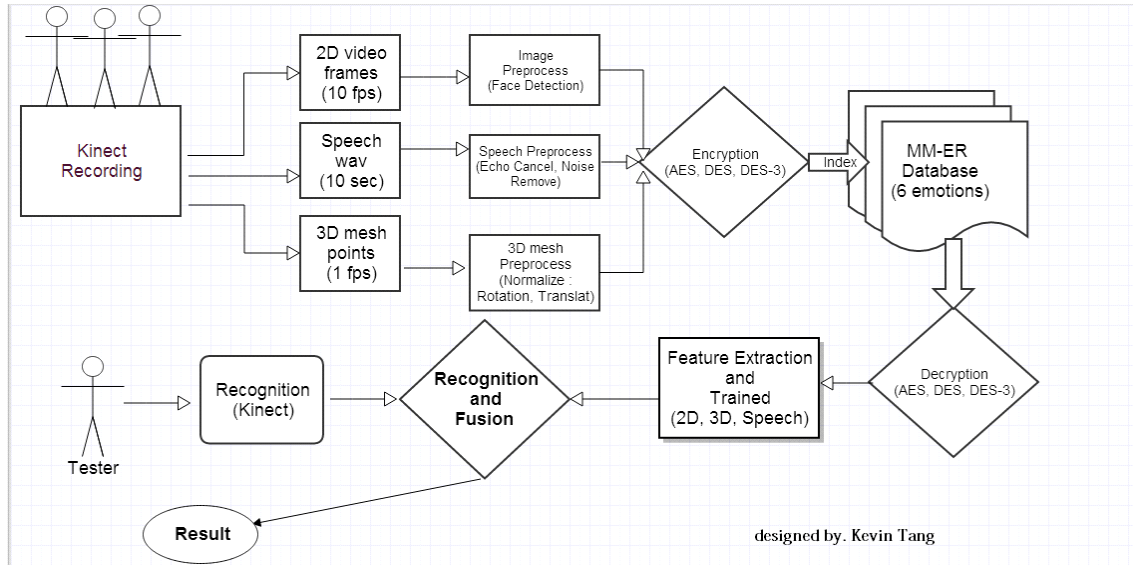


Figure 2.1 Project outline.

hacking on the database, we propose a new information fusion based approach to emotion recognition combined with security features. The following sections briefly explains the proposed system, the challenges that we face in the research, and the contributions of this work.

2.1 The Proposed System

In this section, we given an overview of the proposed system. As shown in Figure 2.1, the proposed system consists of five major parts.

The first part is data collection, which involves using our selected device to concurrently collect the raw data of both 2D video frames, 3D feature points, and audio data. It is elaborated in Chapter 3 - data collection. The second part is pre-processing, of which the main target is to normalize the raw data. This is discussed in Chapter 3 - data pre-processing, which the normalized data than be used to perform the feature extraction. The third part is the secured database to which the raw data will be sent after an encryption procedure using AES, DES and 3DES algorithms. This is discussed in chapter 4 for the data security. Once the data is needed, it can be decrypted

and retrieved. The fourth part is the feature extraction. Feature extraction is important in any recognition system. For example, in order to identify the difference between a banana and an apple, we would compare their shape as well as their colour in the form of numerical values. After that, training is needed to average out features, such as the shape and the colour features in respect to a particular item. In the case of an apple, the trained result should be the circular shape and red colour, which is compared with our target. In this thesis, we use PCA, LDA and CCA for the feature extraction and fusion in the two categories: feature level fusion (FLF) and decision level fusion (DLF). This leads to our proposed fusion method as our three fusion categories to complete the recognition result which is discussed in Chapter 3 and Chapter 6, respectively.

For 2D video frames, red-green-blue-alpha colour modal (RGBA) images are collected, and simple pre-processing is performed, including rotation, constraint, and filtering. The image is then encrypted and stored into the database. The frame rate is set at 10 frames per second and record the video for a period of 10 seconds. Also, face detection and Gabor filter are used to split the facial image into different shape regions, and the closed facial region will be used to separate the face and the background.

For 3D feature points, KINECT provides a face tracking software development kit (SDK) which is able to track the human face, and numerically output the dimensional position in terms of (x, y, z) coordinates. With primary and secondary tracking points, the KINECT SDK is able to track 121 points at once [8]. All 121 points have been used.

For the audio data, there are many methods for audio recognition. The recognition accuracy for audio based identity recognition is now around 99.8%. However, the accuracy of audio-based emotional recognition is only around 65% to 70% [13, 14]. Since the recognition accuracy is relatively low, fusion for multiple modals can improve the overall recognition accuracy. The

research shows that some experiments show better in visual emotion recognition, while the others show that it is better in audio emotion recognition [2]. By consider different modal, it also convey more information about human emotion. More than that, the research show that fusion between multiple modalities on different emotion generally achieve a relatively higher recognition accuracy [6]. The three major features in audio-based emotional recognitions are pitch, energy, and frequencies [3]. Audio wave files are passive and have been added on to increase the dimensionality and accuracy of our fusion recognition system. Audio recording is synchronized with the recording of 2D video frames and 3D feature points. Users can say anything they want for between 5 to 10 seconds.

The characteristic of each dimension can be expressed as follows: the 2D video frame images come from the video. It focuses on the expression of colour in the face as well as the position of fiducially important points, plus the audio part. Based on the psychological research on general emotion types in the past [5, 6], we collected the six emotions outlined in Table 2.1 (fear, happiness, anger, surprise, sadness, and disgust) plus the “normal” or neutral emotion, for a total of seven emotions.

Table 2.1 Characteristics of each data type.

Emotional expression	2D image (facial)	3D points (121 Features)	Audio (10 sec)
Fear	Eyes widely open, mouth may be open	Eyebrow position goes into a reverse semicircle	Volume high or low.
Happiness	Mouth curved up, eyebrows curved down	Mouth line up for upper semicircle	Pace is fast, volume is a little high (energy high)
Anger	Eyes wide open, face turns red	Eyebrow position with a opposite position	Volume is high
Surprise	Mouth open, eyebrow position	Mouth open wide and eyebrow move up	Pace is fast
Sadness	Face turns red	Eyebrow move down mouth close	Pace is slow, volume is low
Disgust	Mouth wide open and AU points of cheek moving toward outside	Mouth wide open, tone position	Pace is slow, volume is low

2.2 Contributions

Multimodal recognition system involves multiple modalities based on different categories of features, including 2D video image frame, 3D feature point and audio data in this work. Our contributions among them include are summarized as follows.

2D video frame images, 3D feature points, and audio data. First, the emotion database with 2D video frame images, 3D feature points, and audio data is the foundation of our work. To our knowledge, this emotion database is the first containing 2D images, 3D vertices, and audio data, with all three types of data collected during the same period of time. This lays the ground for accuracy in fusion-based emotion recognition and potentially reflects a person’s real emotion. The

2D video frames have a resolution of 640×480 at recording rate of 10 fps. The 3D feature points have 121 points of major features at one frame per second (the frames rate can be modified). The audio is a 10-second standard 44100 Hz wave, designed to be language-based independent. Therefore, we have collected emotion in several languages, including English, Chinese (Mandarin), Chinese (Cantonese), Hindi, Punjabi, Korean, Serbian, Russian and Egyptian Arabic.

Modified security database. Our database has been tested using the three general security algorithms, which are AES, DES and 3DES. In the recent cyber-crime cases, the hacker is able to collect more than 1.2 billion username and password combinations and more than 500 million Email addresses. The vulnerable database leads to the risk of lowering the efficiency and reliability of the recognition system [33]. To provide a secure emotion recognition environment, the database needs to be protected. Therefore, these three commonly used encryption standards have been evaluated to test their performance. Based on the processing speed of encryption and decryption, security level, accuracy, and extensible features, we develop an AES database for high efficiency and wide extensibility.

Multi-level fusion method. This method mixed fusion at feature level and decision level. Generally speaking, there are three main fusion categories: data/feature-level fusion, score-level fusion, and decision-level fusion. We further developed a multi-level fusion method that combines decision- and feature-level fusion.

Chapter 3

A Novel Multimodal Emotion Recognition Framework

The proposed multimodal emotion recognition framework is elaborated in this chapter, which includes the data collection, data pre-processing, feature extraction and recognition.

To build a successful pattern recognition framework, a database of sufficient variety is necessary. To the best of our knowledge, there has not been a single emotion database that includes 2D visual information, 3D depth information, and audio for the purpose of the research on multimodal emotion recognition so far. Moreover, there are very few 3D databases for this goal, most of which are developed for face recognition only. To address the dearth of such a key issue in developing a multimodal emotion recognition framework, we collect facial expression and the accompanying audio simultaneously, where the former consists of 2D images and 3D vertices. Two critical issues associated with data acquisition, i.e. the device selection and the recording procedure, are described in details later.

Due to the photometric variation resulting from the change in the light source and the background colour, the collected 2D images need to be normalized in a pre-processing step. Meanwhile, appropriate normalization needs to be applied to the 3D vertices as well as the audio data. For the 3D vertices, since the size of the face varies from one user to another and their faces may be posed at different angles with respect to the camera, the normalization is mainly focused on shifting, scaling, and rotation of the 3D vertex points. As for the pre-processing of the audio data, since each user may start performing at different point of time during the 10 seconds duration,

therefore, to cut the leading and lagging edge which have no relationship to the emotions, this become part of the normalization on audio data, as well as the noise reduction.

Following the data acquisition and pre-processing, feature extraction is the next integral step of a recognition framework, converting heterogeneous data into a common vector-based representation encoding the discriminative information for training a model and classifying new examples. The proposed multi-level multimodal emotion recognition framework is employed to effectively utilize all of the three types of data, and the superiority of its performance to those of the existing approaches is justified in the chapter of the experimental results.

3.1 Data Collection

Data with higher quality can potentially be used to improve the performance of a pattern recognition system. In the context of the presented study, the quality of the collected data is determined by the resolution of the RGB images, the sampling rate of the audio files, and 121 of vertex points in the 3D data. To maintain a certain level of the quality, we are faced with two critical issues, i.e. the device selection and the recording procedure.

3.1.1 Device Selection

Conventional 3D image/video capturing devices, such as Canon XF100/XF105 and Fujifilm Point-and-Shoot, are not capable of producing depth information, and therefore are not under consideration. In recent years, numerous 3D content (data) capturing devices have been developed. The devices generally can be classified into three major categories: the 3D stereo camera released by Gløckner in 2013, the 3D face imaging system released by 3DMD in 2014, and the IR camera released by Qasem in 2012. The most common 3D



Figure 3.1 Kinect.

content recording device is the 3D stereo camera, which uses algorithms very similar to the mechanism of the human visual system based on positional differences that can be used to analyze depth. In contrast to the 3D stereo camera, the 3D face imaging system uses random light pattern projection in the speckle projection flash environment, such as Osela's pattern projection. It can track from 4,400 to 100,000 dots using three cameras and one projector. Finally, the infrared (IR) camera projects the IR into a range of space. Once the IR hits an object, it bounces back to the IR receiver, which can be used to measure the position of 3D content spatially

Taking into account both technological and financial needs of the general public, Microsoft announced a sensor based recording camera, called KINECT, as shown in Figure 3.1 [53]. Similar to KINECT, ASUS also presents an IR camera – XTON. Based on our further investigation of the design of KINECT, we choose KINECT to be the data acquisition instrument. The specific reasons are described below.

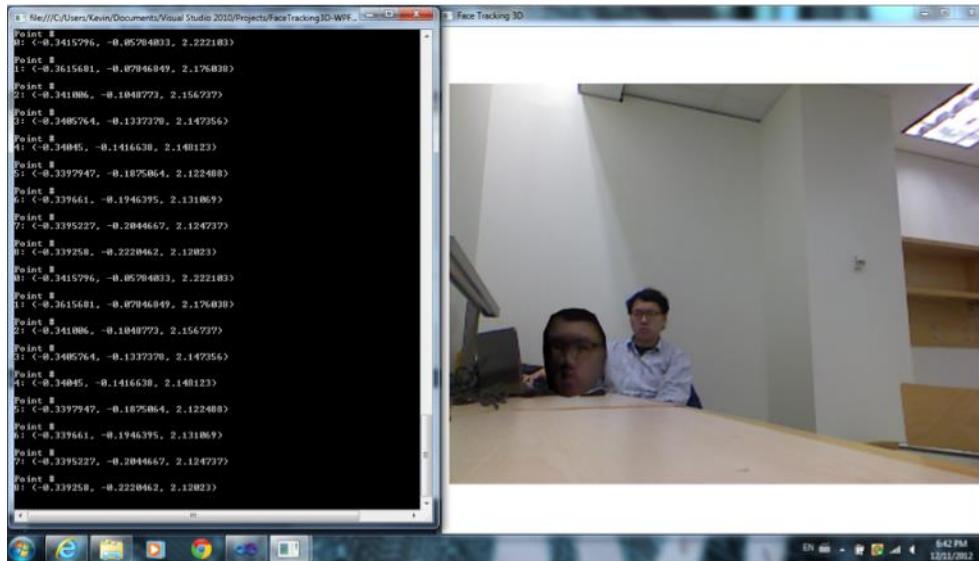


Figure 3.2 KINECT IR camera data.

Similar to conventional imaging devices, KINECT has a colour camera located in the centre, which captures an object's red, green, blue, and alpha (RGBA) colour information stream, each having an 8-bit colour depth. In contrast to conventional equipment in terms of the suitability to our study, KINECT has a number of advantages over the alternatives. There is an IR camera located on the left side of the KINECT and a projector located on the right side. The IR camera captures the object's depth (location) information with a range between 1.2m and 3.5m, and it can show the actual 3D points, as shown in Figure 3.2.

It can capture a horizontal view angle of 57° , shown in Figure 3.3, and a vertical range of 43° . With the assistance of a built-in motor, the camera can move across a range of 27° vertically, as shown in Figure 3.4, in order to provide a wider range of information collection. Within this range, KINECT can collect up to 60,000 vertices points in a 3D

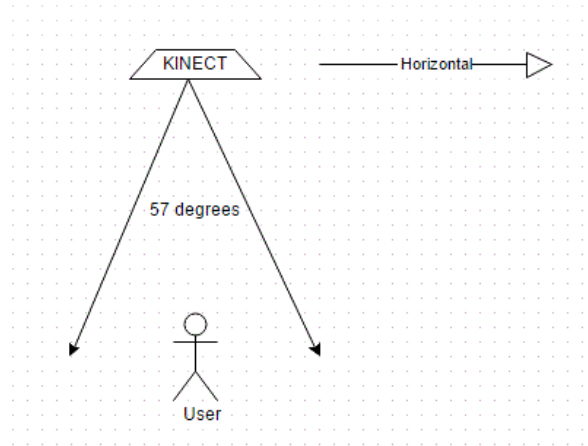


Figure 3.3 KINECT horizontal view angle range.

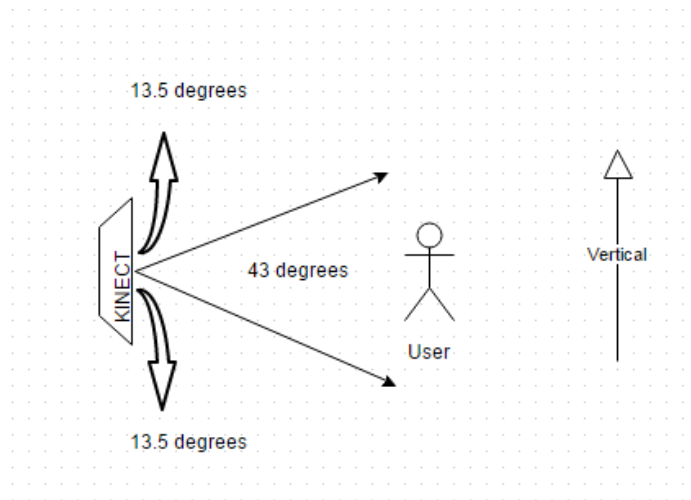


Figure 3.4 KINECT vertical view angle range.

space. As mention before, there is a device similar to KINECT in terms of functionality and size, e.g. the ASUS XTON. However, unlike KINECT, no open source application programming interfaces (APIs) are available for ASUS XTON to generate the data usable in our study.

Secondly, KINECT also has an embedded array of four microphones at the bottom. According to Microsoft, the microphones can detect the differences between volumes of the received voices, which in turn can be used to estimate where an object is located, and predict if it is approaching or moving away. Combined with audio echo cancellation (AEC), it can alleviate the positive feedback in the microphone device.

3.1.2 The Recording Procedure

To collect the raw data in a proper set-up and under an unbiased condition, the general environment factors, such as light source, noise, and other possible interferences, and the condition between each data collecting need to be consistent. Likewise, the human factors, e.g. laughing during the recording of sad emotion, must also be considered in order to avoid and eliminate biased data. To this end, we established a raw data recording procedure, which is summarized in Table 3.1.

Table 3.1 The schematic description of recording procedure.

```
create thread-1
    2d video frame capture
        if fail or bias condition happened
            then exit
create thread-2
    3d content (fiducial points) tracking and capture
        if fail or bias condition happened
            then exit
create thread-3
    audio recording
        if fail of collecting or bias condition happened
            then exit
if mutex not lock
    ---> join thread-1
    ---> join thread-2
    ---> join thread-3
else
    ---> exit
```

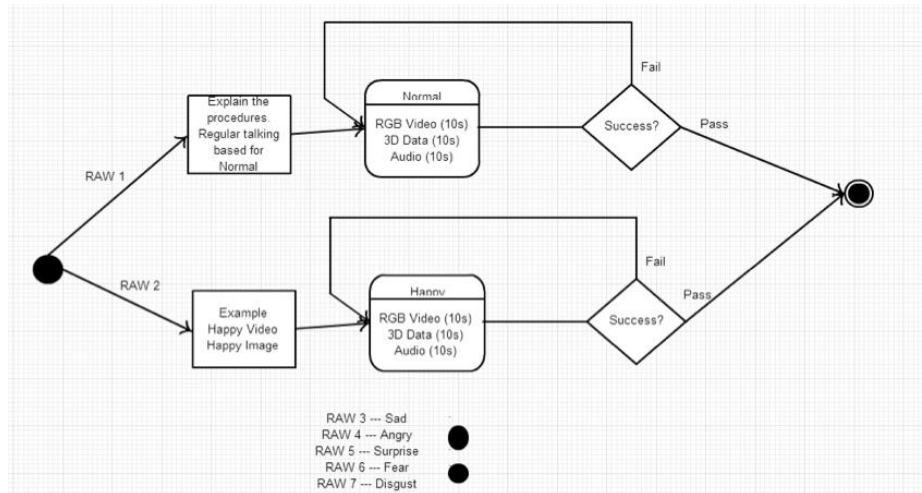


Figure 3.5 Collection procedure.

To ensure the quality of the raw data, the recording procedure shown in Figure 3.5 has been adopted for each recording. As part of the recording process, a subject is required to sign a consent agreement to authorize the collection and the use of the data from his/her emotion information across 2D face images, 3D face feature points and audio voice recording. In addition, an instruction guide is provided on how the 7 emotions, including neutral emotion, should be taken.

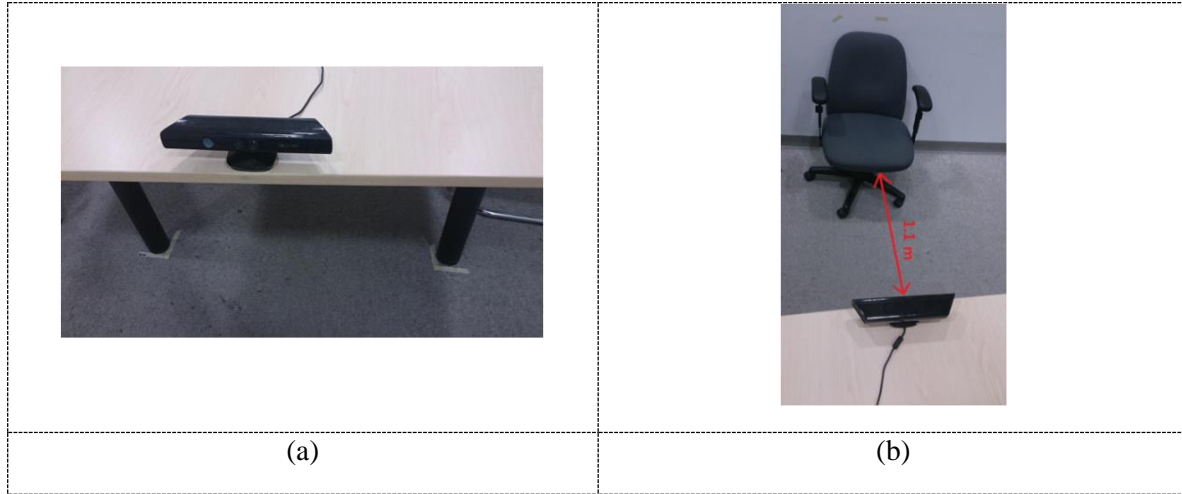


Figure 3.6 (a) Device set-up position. (b) chair position.

A. The spatial arrangement of the selected device

To record the raw data, we first place the KINECT device on the table, as shown in the Figure 3.6. The table is 0.95 meter high and 1.1 meter away from the chair that the camera faces, the room size does not affect the recording quality, but the direct distance between the KINECT and the chair will affect the result. Therefore, 0.95 meter high and 1.1 meter to the chair is the suggested position. This ensures that our basic space coordinates between the device and the users are properly fixed.

B. Light Control and Noise Reduction

The light source in the room also affects the recorded image quality. To ensure every single clip has been recorded under a consistent light source condition that all 12 light bulbs are in neutral condition, they have been replaced at the same time. Each light bulb has been set at the same colour temperature and colour rendering index (CRI) level to 4,000k with 60 Watts, which is white light

instead of yellow light. This is close to a cool white colour, commonly used in an office environment. In addition, since each video clips includes an audio recording part, the noise needs to be reduced during recording. To this end, we consider both internal and external noise. Our selected room has a soundproof interior wall, which reduces most external noise. For the internal noise, we utilize the embedded noise cancellation feature of KINECT since its microphone array is able to isolate the echo using algorithms, such as independent component analysis.

3.2 Pre-processing

Since the characteristics of each individual type of facial expression are primarily associated with its dynamics over time, RGB video frames are collected as the 2D images in our study. The pixel resolution is chosen to be 640x320 to ensure the quality of the frames. In terms of the frame rate for video recording, research shows [41][42][43][44] that human facial expression changes every 0.1 seconds. As such, setting the frame rate to 10 frames per second instead of 30 frames per second helps balance the performance of the system and the redundancy of the data. With 10 frames per second, we take a video clip of 10 seconds' duration per example, resulting in 100 frames of each emotion, as shown in Figure 3.7.



Figure 3.7 2D image data in database.

Prior to storing the collected video frames in a database, they are pre-processed in order to facilitate the learning of the proposed model. To be specific, the video frames need to undergo such procedures as noise reduction and background removal. To this end, we employ Gabor filters to remove the background and subsequently apply a median filter to remove possible noise or dark pixels in each frame. These pre-processed frames are the examples from which the features are extracted at a later stage. An example of the results of the pre-processing for face detection is illustrated in Figure 3.8. Examples of collected data are shown in Figure 3.9.

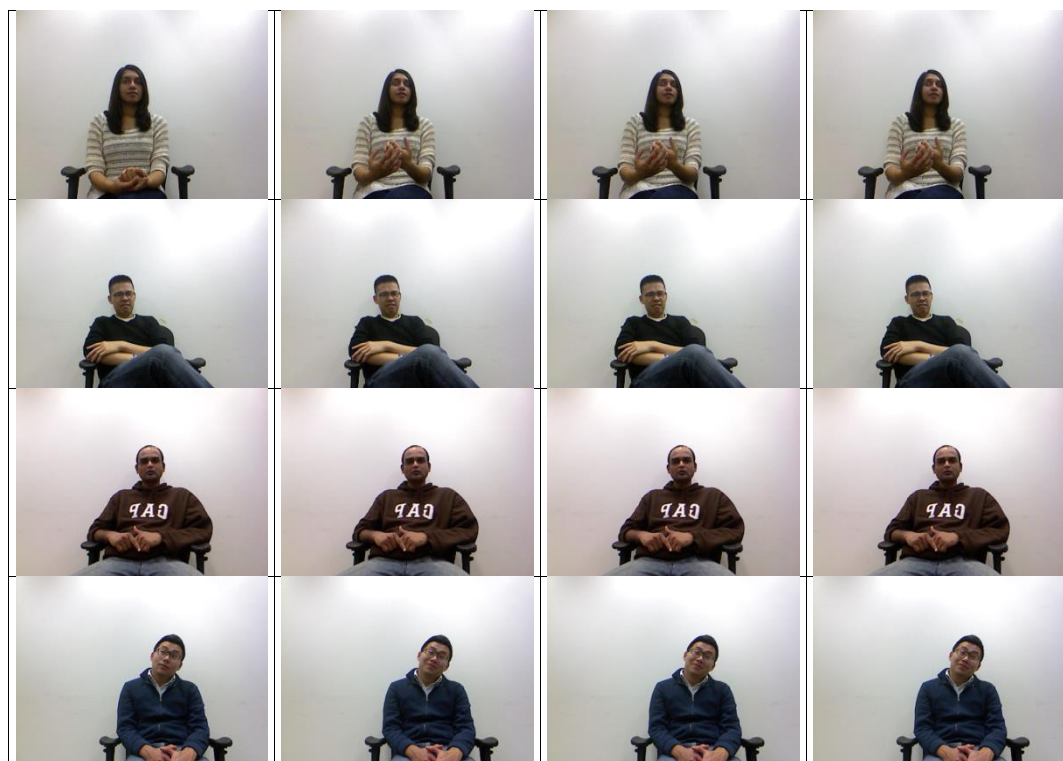
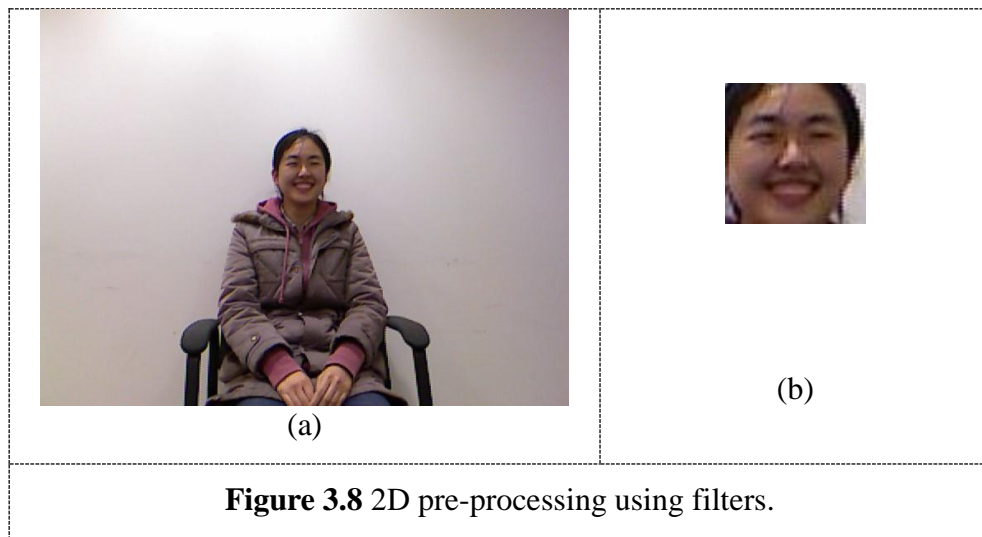
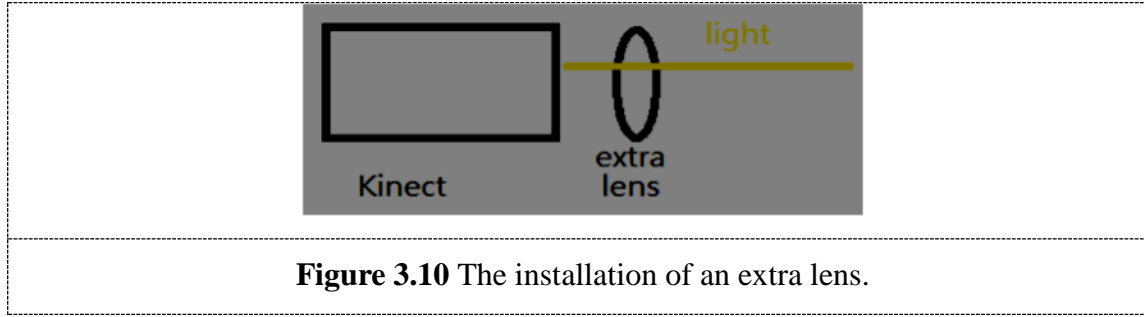


Figure 3.9 Emotional state of disgust.



Improvement on the focus of the 3D data

Because the KINECT sensor has a range limited to 1.4 meters from the device, the subject might not be completely captured. For example, their faces might not be in the range. This immediately leads to the coarse depth resolution and the small number of 3D vertices points collected, which characterizes the capability of a depth sensor in terms of the minimum differentiable distance. The quality of 3D depth data, expressed with a higher depth resolution, is equivalent to finer depth difference, which can help increase the accuracy of the subsequent recognition. To address this problem, a small concave lens is added in front of the IR transmitter, as illustrated in Figure 3.10.

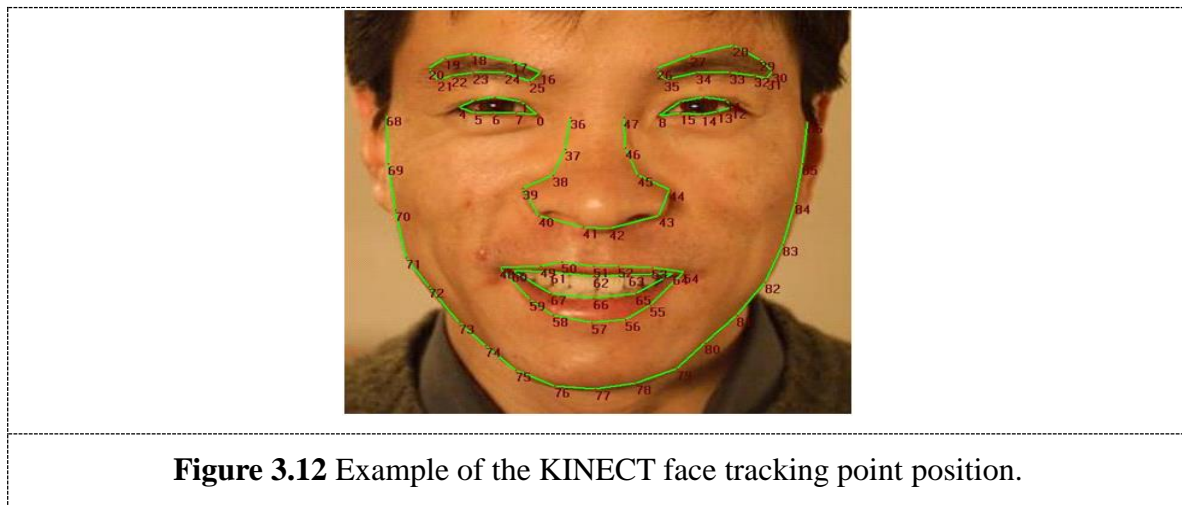
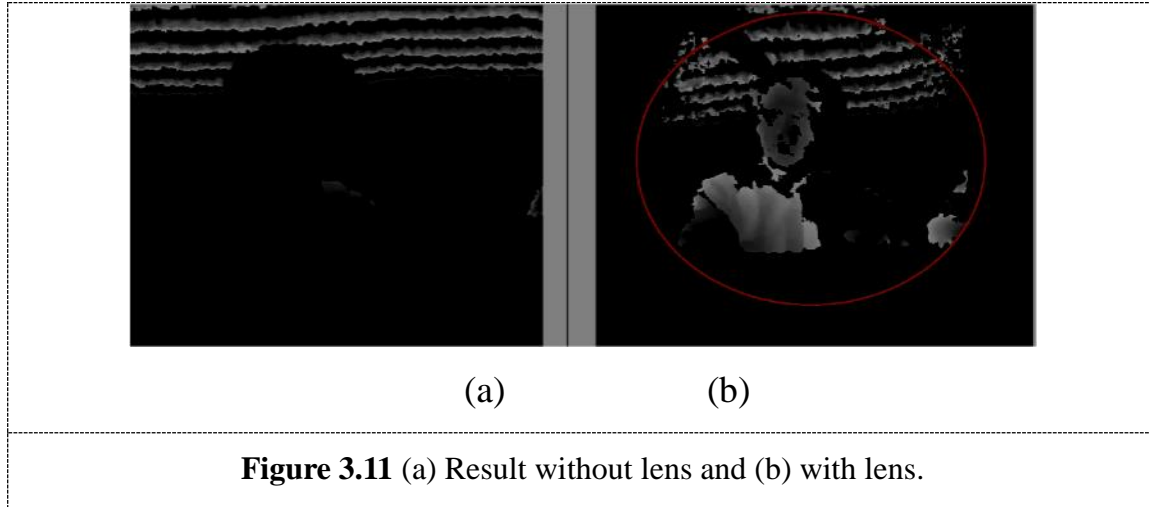
Our object distance can be calculated easily based on the distance between the lens and the KINECT. Based on thin lens equation (3.1), the inverse of the focal length f is proportional to the sum of the inverse of the object distance d_o and the inverse of the image distance d_i ,

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i} . \quad (3.1)$$

When using the lens, we can see more layers in the face region, where each layer represents a depth level. Although we still lose the number of layers on the outer range, as we can see on the left part, we observe clear lines on the top part. But after we install the lens, the lines only appear in the middle part. This is because the lens changes the light path, increasing the number of the layers while decreasing the size of the range. To solve this problem, the subject is required to sit in a position such that his/her face will always appear in a circular area in the middle of the image. The difference between using a lens and without using it can be observed by comparing Figure 3.11(a) and Figure 3.11(b). In order to calibrate the difference between with and without lens, we first place a box in front of the Kinect, by measure the four edge points, we get position as $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$, $P_3(x_3, y_3, z_3)$ and $P_4(x_4, y_4, z_4)$. Next, we install the lens, but not moving either box or Kinect, and we measure the same edge points, we get $P_1'(x_1', y_1', z_1')$, $P_2'(x_2', y_2', z_2')$, $P_3'(x_3', y_3', z_3')$ and $P_4'(x_4', y_4', z_4')$. By moving the box to different place, we collect another 5 sets of total 40 points (20 with lens, 20 without lens). By using matrix as following, we can then calculate the shifting and scaling operation constant for k_x , k_y and k_z as well as b_x , b_y and b_z as following.

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} k_x & 0 & 0 \\ 0 & k_y & 0 \\ 0 & 0 & k_z \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix}. \quad (3.2)$$

The advantages of KINECT sensor include not only its price, but also the associated open source software for image processing. The KINECT Face Tracking SDK is able to track 121 feature points in 3D data on a detected face. The positions of the points on the reference picture from Microsoft are shown in Figure 3.12 [20]. Note that not all 121 points are shown here because



some of the points are secondary points, which are not related to emotion information. Hence, we would not use those points for recognition.

After collecting the 3D data set, we store it in the same order as the FaceTracking SDK (Microsoft) does, and the detailed position information is then stored into a single text file. There are 10 sets of 3D content data during the 10-second recording period. According to our experiments, every person may take a different pose during the recording, and the face position is changing from

time to time. Moreover, during the 10-second period, not all 10 frames (or more frames) represent the desired emotion. Therefore, we need to perform pre-processing for the 3D points data. The 3D data pre-processing includes two major steps: dimensionality reduction and normalization. For the dimensionality reduction, we use OpenSceneGraph to display the 3D points, build the 3D mesh, and remove certain frames of secondary importance to reduce the number of frames. If a frame have no significant difference or the same as the pervious frame, than we consider it's as the secondary importance frame, which can be removed away. In normalization, as mentioned before, the face mesh may not be set at the same position every time, and therefore, the 3D data need to be normalized in order for comparison. To normalize the data, we set the noise tip point to the origin $(0, 0, 0)$. The reason why we choose this point is that it has the smallest Z axis value, which is the closest point to the camera, by shifting all of the 121 points with the same vector as we move the noise tip points. As shown in equation (3.3), we locate the position of noise tip denoted P, of which the coordinate is represented using (x_p, y_p, z_p) . By calculating the vector between noise tip to the centre $(0, 0, 0)$, we get the direction vector (x_p, y_p, z_p) , i.e.

$$V(x_p, y_p, z_p) = P(x_p, y_p, z_p) - (0, 0, 0). \quad (3.3)$$

Next, we apply matrix subtraction between the original position P_{121} and the direction vector V , that is

$$P'_{121} \begin{bmatrix} x'_1 & \cdots & z'_1 \\ \vdots & \ddots & \vdots \\ x'_{121} & \cdots & z'_{121} \end{bmatrix} = P_{121} \begin{bmatrix} x_1 & \cdots & z_1 \\ \vdots & \ddots & \vdots \\ x_{121} & \cdots & z_{121} \end{bmatrix} - V \begin{bmatrix} x_p & \cdots & z_p \\ \vdots & \ddots & \vdots \\ x_p & \cdots & z_p \end{bmatrix} \quad (3.4)$$

which results in the final P'_{121} for all 121 points. Then, this new set of points is referred to as P'_{121} .

The next step is to consider XY plane rotation, where the XY plane is the horizontal plane. There is a central vertical line across a face connecting the middle point between the eyebrows and the middle point of the mouth. Let us denote this line L_{XY} based on human face geometry. This line

should be vertically straight, which means it will be parallel to the Y-axis. Denote the angle between L_{XY} and Y-axis as θ' , any point that stays on the right-hand side of L_{XY} will be rotated by θ' clockwise, and any point on the left-hand side of L_{XY} will be rotated by θ' counter-clockwise, where θ' can be calculate by

$$\theta' = \begin{cases} \emptyset & \text{if } L_{xy} \text{ on the positive side,} \\ -\emptyset & \text{if } L_{xy} \text{ on the negative side.} \end{cases} \quad (3.5)$$

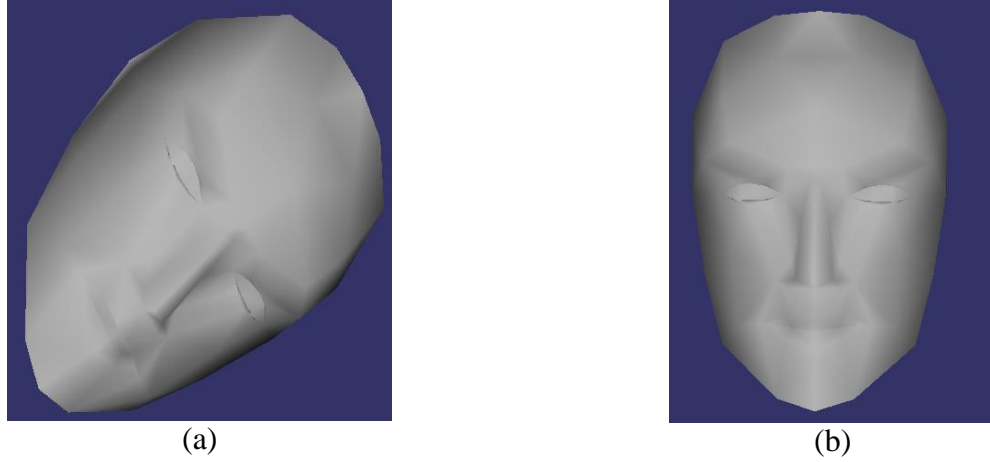


Figure 3.13 (a) before 3D pre-process (b) after 3D pre-process.

The rotation for XY plane can be formulated as equation (3.6), where (x, y) is the starting point position. After rotation, only the x and y of the coordinates of the points will change, the z component remains constant. Based on matrix rotation transformation with respect to θ , we get the new coordinate (x', y') of a point by

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta' & -\sin\theta' \\ \sin\theta' & \cos\theta' \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \quad (3.6)$$

Then the last one is the rotation in the YZ plane. Notice YZ plane is the vertical plane. After the vertical rotation is done, all 121 points are transformed to the normalized positions. However, although we normalize the data, the rotation angles will be recorded and stored into the



Figure 3.14 Pre-process for emotion audio.

data, because this information can provide related information for emotion expression. The process is shown in Figure 3.14.

Audio recognition is different from 2D image recognition and 3D content recognition because the content of audio cannot be visualized. Therefore, in audio recognition, we focus on the tones, the volume, and the pace of the audio. The waveform needs to be investigated in both the time domain as well as the frequency domain. The KINECT sensor has an array of embedded four microphones, which is able to locate a speaker's position by comparing the volume difference between the microphones. Each sample of the audio data we collect is 10 seconds in length and sampled at 256KB/sec, which provides a sufficient quality for audio emotion recognition. The pre-processing of audio data focuses on three things. First, it removes unrelated information. Although each recording session is 10 seconds long, the actual temporal duration for emotion expression is on average less than 4.75 seconds. This means the other 5.25 seconds do not contain important emotional audio information, and hence can be removed. Second, noise is removed. There are two types of noise: the stationary noise and non-stationary noise.

In our recording environment, the noise coming from the computer devices and the air conditioner belongs to the stationary noise. Therefore, as shown in Figure 3.14, the first thing for

pre-processing is the noise reduction. There are multiple methods to perform noise reduction. We select a wavelet-based method because of its efficiency and its advantage of preserving the original audio signal. Afterwards, we remove the leading edge. This includes the preparation time for the collector to enter the actual moment of emotion expression. Finally, we remove the trailing edge, which includes the silent moment after the emotion has been expressed. With all three pre-processing steps done, useable emotion data can be stored into a database.

One of the key concerns in audio emotion recognition is that it is typically different than visual emotional expression. Since different languages have different sound, our target is to design a language independent audio emotion recognition system. In the audio data collection, we also consider the language effect factor, as shown in Figure 3.15 [48]. Human language can be grouped into families of similar languages. For example, both Cantonese and Mandarin are in the family, shown in red on the map in Figure 3.15. It will be very different compared to Hindi, which is shown in green on the map. On the other hand, Hindi is not too different compared with the light green of western European languages, such as French and English. We use this standard to list the categories that the language families belong to. Based on the research on human language behaviour, we define that similar language families will have similar behaviours, such as tone and response.

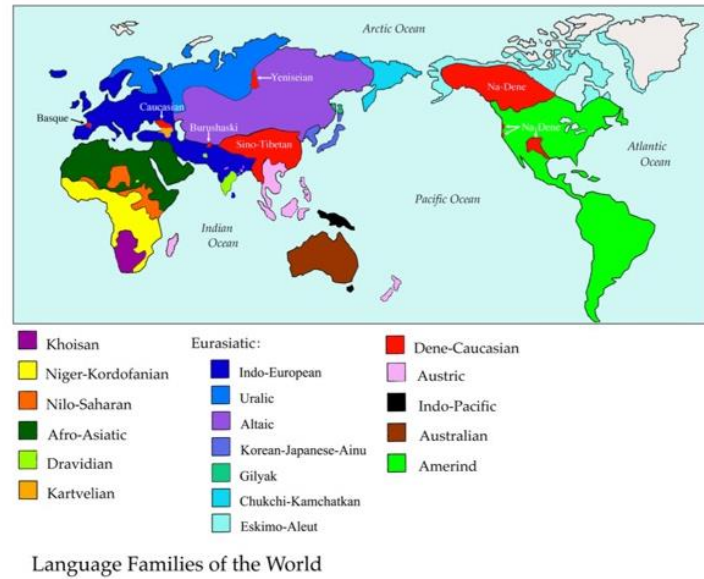


Figure 3.15 Language relationship.

Shown in Table 3.2 is the size of our database for the audio data part. During the recording, we ask each volunteer to use his/her primary language, which is not necessarily English; however, we do not force them to speak out with every single emotion because the primary goal is to collect samples of spontaneous emotion. Therefore, asking the volunteers to collect the data under a relax environment and procedure is important.

Table 3.2 Languages analysis on audio data.

Index	List of languages	Number of RAW audio data	Language categories ^{L1}	Pre-processing
1	English	23	Dark Blue	22
2	Chinese (Mandarin)	36	Red	36
3	Chinese (Cantonese)	12	Red	11
4	Hindi	15	Blue	12
5	Punjabi	6	Blue	6
6	Korean	6	Light Blue	6
7	Persian	6	Blue	6
8	Russian	21	Light Blue	21
9	Egyptian Arabic	6	Dark Green	6

As shown in the Figure 3.16, the two spectra depict the temporal/frequency patterns of Korean and Persian language when the happiness emotion is being experienced. This provide a variety languages based audio database.

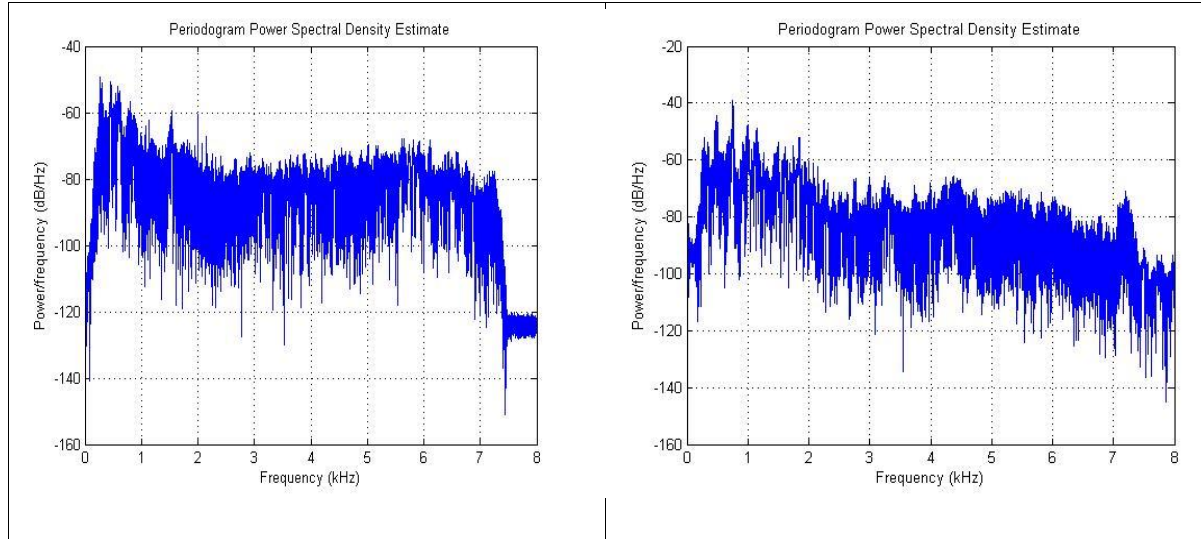


Figure 3.16 Comparison between Korean (left) with Persian (right) languages.

3.3 Feature Extraction

Feature extraction for the multimodal facial expression recognition system treats 2D, 3D and audio differently due to their distinct characteristics, which means each modality needs its own feature extraction algorithm. Our primary target of this study is not to focus on the performance of each individual model. Instead, we consider all three modalities. Collecting more feature information may lead to a lower recognition accuracy because of the increased data dimensionality, which also increases the processing time.

To recognize the emotion based on the audio data, the automatic speech recognition (ASR) is needed. Although ASR is mainly used for user identification by audio, the research of emotion recognition by audio has received more attention recently. The first step is to detect the audio, and the second one is to perform the audio feature extraction. The most commonly known feature for audio is the MFCCs. When extracting MFCCs, the power spectrum (PS) of the sound is taken

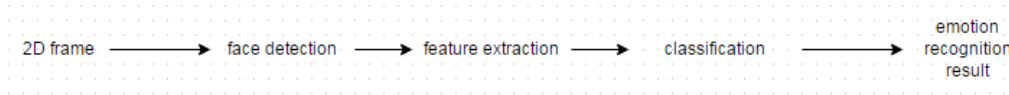


Figure 3.17 2D Feature extraction and recognition process.

to perform the log linear cosine transformation, and then the result considers both Mel scale and frequency scale measured in Hertz (Hz), which is very close to the mechanism of the human auditory system. Currently MFCC is widely applied to general audio recognition, although MFCCs are weak for high frequency band, since Mel-scale becomes relatively inaccurate when the frequency goes up [21].

Furthermore, the MFCCs can be applied to more detailed audio features. According to our best knowledge, the following important audio features are mainly used in audio based emotion recognition. They are excitation source features, prosodic features, and vocal tract features [2]. Each of the features has its own characteristics and field of application. We will discuss and examine them in more details later in this chapter. On the other hand, the 2D image feature can be the feature point location, feature point colour, curving level between important regions. When designed and combined properly, the resulting feature can increase the recognition accuracy as well as reduce the processing time. This concept is similar to audio based facial expression recognition. We do not use a single feature, but combine multiple features for audio feature extraction, such as the prosodic, MFCCs and formant frequency. One of the example from I. Patel and Y.S. Rao's research using MFCC features in combination with prosodic, MFCCs and formant frequency feature results in around 68% recognition accuracy for 6 standard emotions [21].

3.3.1 2D image based feature extraction

To perform 2D emotion recognition, the first step is to perform feature extraction, and then compare the features with a trained model to perform emotion recognition, as shown in Figure 3.17.

As discussed earlier in this chapter, the current methods mostly focus on features on the veins and shape appearance on the face. Although this provides an accurate result, in a multimodal recognition system, the feature information on veins and shape can also be collected from 3D data. Although the resolution of 3D sensor is usually lower than that of a 2D camera, 3D data provide the depth information which is not included in 2D data. This leads to the overlap on information and reduces the processing time. To address this problem, we take into consideration colour features in 2D data. The KINECT camera collects RGBA colour information using a complementary metal–oxide–semiconductor (CMOS) camera. The colour data is one of the features in face-based emotion recognition. In RGBA, the R represents the red colour intensity of a pixel, the G represents the green colour intensity of the same pixel, and the B the blue colour intensity of the same pixel. Finally, the A, i.e. the Alpha, is the opacity channel. If the value of alpha is equal to 0, the pixel point is fully transparent. On the other hand, when the value of alpha is equal to 100, it is a fully opaque pixel point. However, compared with RGBA colour model, the hue-saturation-value (HSV) colour model has been commonly used in computer graphics. The reason is that the HSV is more consistent with the human colour perception than the RGBA. The HSV colour model is shown in Figure 3.18 [35]. In order to convert from RGBA to HSV, we used the following equations. First, the initial hue value H_i is calculated by

$$H_i = \cos^{-1} \left\{ \frac{\frac{1}{2} [(R - G) + (R - B)]}{\sqrt{(R - G)(R - G) + (R - B)(G - B)}} \right\}. \quad (3.7)$$

After H_i is calculated, the intensity values of blue and green are compared. If the intensity value of blue is greater than or equal to the intensity value of green, we set the

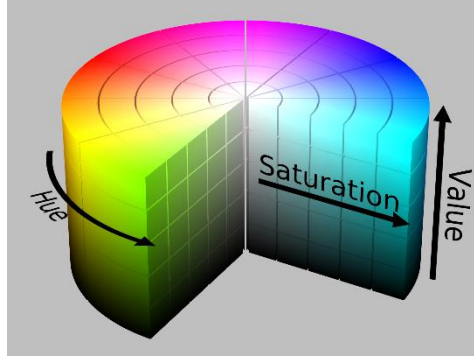


Figure 3.18 HSV colour modal.

value of hue H to H_i . Otherwise, H is equal to 360 degree subtracting the value of H_i as in

$$H = \begin{cases} H_i, & \text{if } B \leq G \\ 360^\circ - H_i, & \text{if } B > G \end{cases} \quad (3.8)$$

The saturation is simply calculated through

$$S = \frac{\max(R, G, B) - \min(R, G, B)}{\max(R, G, B)}. \quad (3.9)$$

Last, the V, value, can be calculated by

$$V = \frac{\max(R, G, B)}{255}. \quad (3.10)$$

Given that the resolution of facial images is 70x70 pixels in each of the three colour channels, each image can be represented using a 14700-dimensional vector. This amount of data includes only one frame of facial emotion in the 2D video data. Next, in facial recognition, although we usually do not need to consider the motion information, facial expression changes over time. Therefore, the analysis of the feature point motion vector is useful in a facial expression recognition system. To analyze that, we use optical flow, which is one of the well-known methods in dynamic feature analysis. The extraction of optical flow can be expressed as

$$v(d) = v(dx, dy) = \sum_{x=ux-wx}^{ux+wx} \sum_{y=uy-wy}^{uy+wy} (I(x, y) - J(x + dx, y + dy))^2. \quad (3.11)$$

This will take into consideration the motion of feature points, which improves the weakness. We collect 7 main points for extracting the motion vector features around the right eye, left eye, centre of the nose, top edge of mouth, bottom edge of mouth, left edge of mouth and right edge of mouth.

Overall, our 2D facial expression recognition uses three feature components, which are the shape of critical areas (such as mouth), the motion vector and the colour feature in HSV colour model. Each of these features has been collected and processed through PCA and LDA to perform dimensionality reduction. The nearest neighbour (NN) is employed as the classification method.

3.3.2 3D vertices based feature extraction

As in the case of 2D, a 3D facial expression recognition procedure includes feature extraction and classifier training to perform recognition. The classifier used here is also NN. In our work, we track

the important feature points based on our experiment, the selection concept also considers the theory of Facial Action Coding System (FACS). The KINECT face tracking SDK is able to track 121 points. They are redundant and contain feature points with little discriminative information, which may affect the recognition accuracy negatively. Therefore, instead of using all 121 points, we select 32 among them; this selection is based on the primary points, i.e. the nose tip, the eye centre points. These 32 points also correspond to the 2D positions of the 2D feature extraction points. As discussed in the previous chapter, after the 3D normalization of the data, the next step is to train the classifier. The frame rate is set to 10 fps for 3D feature points tracking. Although there are many different algorithms for 3D facial expression recognition, one step is to perform Iterative Close Point (ICP) algorithms. The algorithms can be described below:

Algorithm 3.1: Algorithm for Iterative Close Point

```

1.     $T(i_s) = T, i = 1$ 
2.    Do
3.        For  $T(i)$  where every testing points, which  $1 < i < N$  ( $N = 121$ )
4.        Search the closest point  $a(i)$  where  $1 < j < M$  on  $A(k)$ , which  $1 < k < N$ 
5.        End For
6.        Located closest points  $a(j_1)$  based on distance
7.        the pairs of points  $\{(T_1, A_1), \dots, (T_2^N, A_2^N)\}$ 
8.        describe the correspondences between  $T_1$  and  $A(l)$ .
9.        If error of registration  $e$  between  $A(k)$  and  $T(i)$  is too large
10.       Compute transformation  $T(l)$  between  $(T_2(i), a(i))$ , translation and rotation.
11.       Perform transformation  $T_2(i + 1) = T(i) \cdot a_2(i), i ++$ 
14.       Else
15.           Stop
16.       End If
17.       While  $\#T_2(i + 1) - P_2(l) \rightarrow$  take threshold
18.       End

```

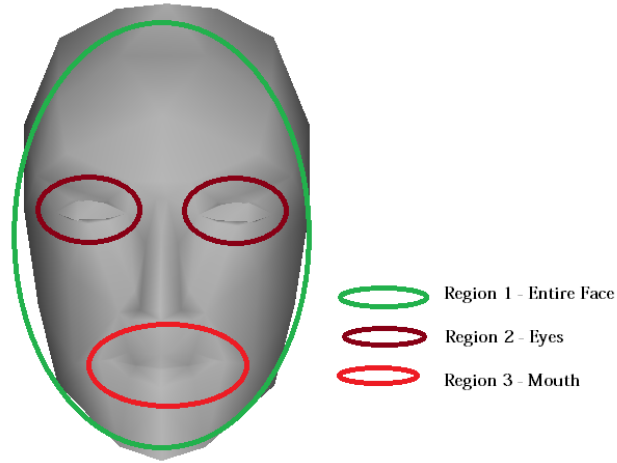


Figure 3.19 Global and local ICP.

Simply, we can conclude ICP as three steps: selecting, matching and weighting. We use 32 primary 3D feature points $P_1(x, y, z)$ to $P_{32}(x, y, z)$ and performed the matching by comparing each of them with our six emotions models plus the neutral emotion. This leads to seven different weighting results, the highest of which is our result. For the purpose on the fusion part that needs to fuse the 3D data with other sets of data, we propose a modified method to use the ICP, which considers both global and local characteristic of facial expressions. As shown in Figure 3.19, we use three major categories: one global part (green part) and two local parts, i.e. the eyes and the mouth, by averaging out each region to get a middle point. The reason we consider these two local parts is because human emotion expression appears more detailed and visually different in these two local parts.

Overall, each vertices point consists of x , y and z parts. Therefore, the total number of features we have in 3D vertices is $124 \times 3 = 372$ features.

3.3.3 Audio based feature extraction

As previous research shows, audio contains approximately 15% of emotion characteristics, which means that performing an emotion recognition without considering the audio component will have at least 15% lower accuracy (the user may not be aware of it). After the pre-processing discussed earlier, the next step is to perform audio feature extraction and recognition. To look at audio-based emotion recognition, we need to first perform the feature extraction. The difference is the audio is not visible compared with 2D images or 3D vertices positions. To extract the feature, we must understand how human audio appears as audio and how this changes when emotions change. According to [15], “The vocalized form of human communication is termed as audio, each of our spoken word is created out of phonetic combination of a limited set of vowel and consonant audio, which are the sound units in audio synthesis” Even speaking the exact same word(s), with different speed, loudness, pitch, and accent, including both cultural and age-related differences, may lead to different results.

Considering all possible details would be extremely unwieldy and difficult to analyze. To simplify the process, we extracted the audio data into MFCCs based on the audio cepstrum which represents a nonlinear spectrum-of-a-spectrum, which is based on Mel-frequency, which appears more similar to what the human ear hears in frequency. In this work, we record 10 seconds of audio using KINECT microphone array, with the data stored as wave audio file (wav), where each file is approximately 250KB to 300KB. The

Table 3.3. Audio features.

Index	Feature Description
1	Audio Length
2	Energy Max (dB)
3	Energy Min (dB)
4	Energy Mean (dB)
5	Pitch Mean
6	Pitch Median
7	Pitch Min
8~20	13 MFCC features

wav file is a standard audio file format which is an uncompressed high quality sound file. This is important because it is able to keep all the original information. So at the feature extraction stage, we can collect more realistic information compared with non-wav files. The audio features we collect are list in Table 3.3.

In MFCC feature, we have the audio signal $a(n)$, after the same pre-process as discussed in Chapter 3, we get pre-processed audio file as $a'(n)$. Next, a window is applied to $a'(n)$, resulting in $a_t(n)$. Then, we perform discrete Fourier transform (DFT) to $a_t(n)$, followed by taking Mel filter banks, to get $m_t(n)$. Then we calculate the logarithm of the absolute value to power of two. Finally, Inverse Discrete Cosine Transform (IDCT) is applied, which results in the MFCC features. Since the audio data contains the duration of 5 seconds per sample, with frequency as 100 samples per seconds, the total number of feature in audio data is $(5 \times 100 \times 13) + 7 = 6507$ features.

3.3.4 Dimensionality reduction

Next, we will discuss the core of the multi-modal face expression system, i.e. the dimensionality reduction. This is the second step in three major procedures of the recognition algorithm.

With each feature in 2D image, 3D vertices and audio data; we perform training to get a trained matrix. As we discussed in the previous sections, the total number of features in 2D image is around 14700 features. This is too large to perform the classification accurately. For such a high dimensional feature space, we need to perform dimensionality reduction. We select PCA as the dimensionality reduction algorithm. The algorithm works as following. First, find the difference between the mean of dimension to each dimension. As shown in equation (3.12), where Ω denotes the entire set of samples in one dimension set, n is the number of samples, Ω_j is the component of set Ω , result with mean of dimension set

$$\bar{\Omega} = \frac{\sum_{j=1}^n \Omega_j}{n}. \quad (3.12)$$

Next, we calculate the covariance matrix. Note the covariance is a method to measure between the dimensions to the mean, to see how the two vary between each other. The covariance matrix can be calculated based on

$$covariance\ matrix = \frac{\sum_{j=1}^n (\Omega_j - \bar{\Omega})(\Omega_j - \bar{\Omega})}{(n-1)}. \quad (3.13)$$

where $\bar{\Omega}$ represents the mean of dimension set. We are able to calculate the eigenvectors of the covariance matrix where the highest eigenvalue being the first principal component. Arrange this data vector by projecting the initial data to the principal component vectors, we can achieve the matrix with eigenvectors V_e . Denote the initial data $\Omega_{initial}$ after the dimensionality reduction, the resultant data denoted Ω_{PCA} , can be computed by

$$\Omega_{PCA} = V_e \times \Omega_{initial}. \quad (3.14)$$

In fact, PCA is an unsupervised machine learning algorithm. In order to improve the quality of dimensionality reduction, we then applied the Linear Discriminant Analysis (LDA). LDA is a typical supervised algorithm, LDA provides the functionality of separate the data groups. However, by looking at the vector norm, the accuracy may not be as good as that of PCA. Therefore, we usually performed LDA after PCA, in order to achieve the best result. Generally speaking, LDA is a way to project data into a higher dimensional space first, followed by performing the dimensional reduction. Assuming the scatter matrices to be S_i , the basis matrix then becomes S_w . We have n classes in this case. LDA is performed according to

$$S_w = \sum_{i=1}^n S_i. \quad (3.15)$$

Next, we find the covariance matrices S_B based on the mean with i_{th} mean in each set, i.e.

$$S_B = \sum_{i=1}^n n_i (m_i - m)(m_i - m)^T. \quad (3.16)$$

Finally, with S_w and S_B it leads to the eigenvector as in

$$S_B W_i = \lambda S_w W_i. \quad (3.17)$$

In order to find w , we calculated the i_{th} column vector, where the i_{th} eigenvalue corresponded to one, which leads to the eigenvector. With the eigenvector, we can use the same method in equation (3.13) to obtain the resultant data.

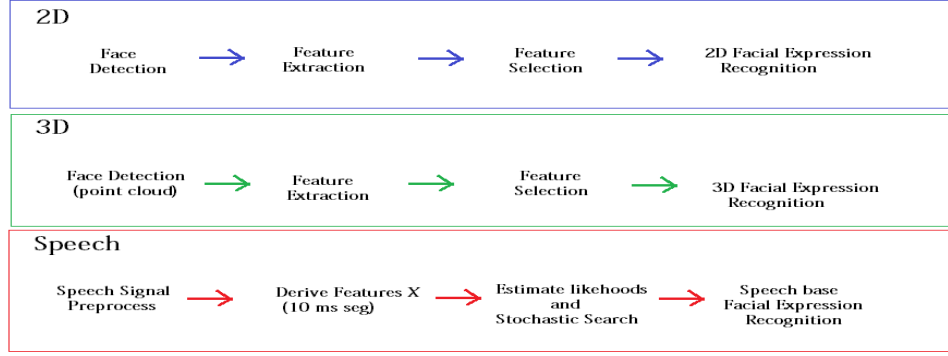


Figure 3.20 Classification process

3.4 Proposed fusion method

The next important step after feature extraction and recognition is fusion. Fusion is the core part of the entire multi-modal recognition system. The process of fusion involves fusing multiple different sets of data by using mathematical algorithms. This process will reduce the complexity of the data sets in order to more quickly and accurately classify the information in the data sets. In this work, the three data sets are 2D images, 3D vertices and speech audio waves as shown in Figure 3.20. In order to combine or fuse them together, we use one of most commonly used fusion methods - the canonical correlation analysis (CCA) for multi-modal data fusion.

Generally speaking, there are two different fusion methods, the decision level fusion and the feature level fusion. First, decision level fusion (DLF), as shown on Figure 3.21, performs the recognition in each data set, including 2D recognition, 3D recognition

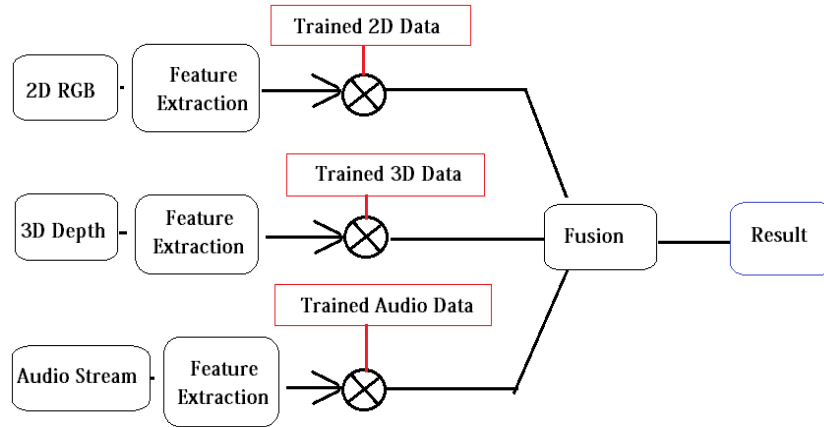


Figure 3.21 Decision level fusion.

and audio recognition. The recognition results will be fused based on fusion algorithms, such as CCA which is used in this work. However, CCA is only able to fuse two different types of data together at a time. Therefore, we fuse 2D with 3D data first, which is in turn fused with audio data to get the final result.

On the other hand, feature level fusion (FLF) is different at the point when the data are joined. As shown in Figure 3.22, the FLF joined all three data types (3D, 2D and audio) into one stream data vector based on the selection of our features in each data. With the joined data, it then performs fusion with the trained features. This leads to the recognition result.

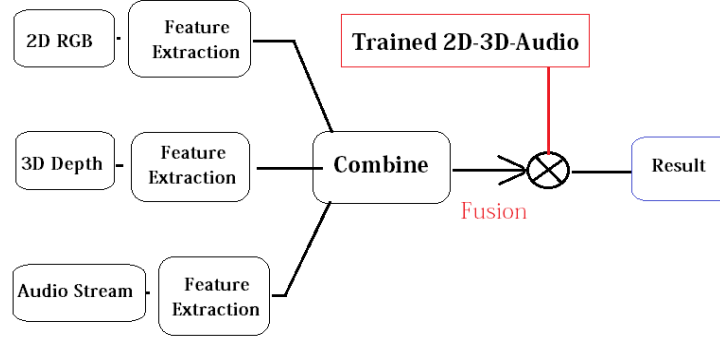


Figure 3.22 Feature level fusion.

FLF is stronger at fusion speed. However, it may not be able to treat every single feature equally, while DLF does. Also, the weighted sum is one of the techniques used here. Equation (3.18) shows how weighted sum works. In weighted sum, the result is taken as R_m , which is equal to the product between the weighted constant value w_m and the number of the data set d_n from the range of 1 to N , i.e.

$$R_m = \sum_{n=1}^N w_m \times d_n. \quad (3.18)$$

Canonical correlation analysis (CCA) can be formulated as follows. We have two sets of variable groups, x and y , where each variable group can form a linear relationship, $x_1^\wedge = a_1'x$ and $y_1^\wedge = b_1'y$ respectively. In x , it contains N variables in the matrix, and in y it contains M variables in the matrix. The result, P , can be described as

$$P(x_1^\wedge, y_1^\wedge) = \frac{Co\varphi(x_1^\wedge, y_1^\wedge)}{\sqrt{Va\varphi(x_1^\wedge) \cdot Va\varphi(y_1^\wedge)}}. \quad (3.19)$$

Our proposed fusion method combined both FLF and DLF into one fusion. By using CCA, we technically perform fusion on a fusion.

Considering DLF as one independent result, it is then fused with the feature level fusion on each data set (2D, 3D and audio). The main issue is that it is more than 2 components. Therefore, we cannot use CCA directly. We need to fuse the features first between the 2D frame image and the audio. Then after that we perform the fusion again with the 3D vertices data. This proposed method can generally be described by equation (3.20), where R represents the final result, the $2D\ feature$ is the resulting 2D image feature recognition, the $3D\ Feature$ is the resulting 3D vertices feature recognition and $audio\ feature$ represents the audio recognition result, i.e.

$$R = \sum(2DFeature \oplus 3DFeature \oplus AudioFeature \oplus \sum(2DFeature + 3D\ Feature + Audio\ Feature)). \quad (3.20)$$

To be clearer, we basically combined both FLF and DLF, but treated DLF as another result. Therefore, all four of these were combined for the fusion.

3.5 Classification

Classification is the final step of the recognition system. There are multiple classification methods in existence. In our work, we first used the nearest neighbor method. The nearest neighbor method is a non-parametric algorithm [49]. Assume we have M amount of classes existing in the recognition database S , and assume our testing example is X . Each class set can be represented as in equation (3.21), where Ω_1 is the first class in the database and Ω_M is the last, that is

$$S = \{\Omega_1, \Omega_2, \Omega_3, \dots, \Omega_M\}. \quad (3.21)$$

Next, we take the trained result in each class. We would also have M trained results represented as

$$Ts = \{T_1, T_2, T_3, \dots, T_M\}. \quad (3.22)$$

Finally, we find the minimum distance using Euclidean distance rule. In order to find the Euclidean distance between any two points, $P_1(x_1, x_2, \dots, x_n)$ and $P_2(y_1, y_2, \dots, y_n)$, in N dimensions, equation (3.23) is used

$$d = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \dots + (y_n - x_n)^2}. \quad (3.23)$$

Based on this, we test example X to each trained result from T_1 to T_M . The result can be calculated by

$$i_r = \arg \min_i d(X, T_i). \quad (3.24)$$

where i_r is the identification of class.

Chapter 4

Database and security

4.1 Database

The database is always the foundation of any type of recognition system. In order to collect the data, a computer vision device is needed—such as the camera and the microphone. In this work, since the data represents the collection of users' facial expression in real time, and consists of 2D images, 3D vertices, and speech simultaneously, the selection of device is important. Also, since the RGB image could be affected by the recording environment (such as the light source in and background colour of the recording room), we need to perform pre-processing to normalize the 2D image, the 3D vertices and the speech audio data. This section discusses collection and pre-processing procedures for 2D image, 3D vertices, and speech data.

Each data set includes a personal information component and a data component. Personal information recorded the gender, race, and age of the user. It includes six emotions (fear, happiness, anger, surprise, sadness, and disgust), and each contained a 100-frame 2D images, 10-set of 3D points text data, and a 10-seconds long speech wave file.

Since the data are directly stored in the computer hard-drive, there is potential security risk related to the database with regards to sabotage or theft; therefore, we added a cipher encryption procedure which is discussed in the next chapter.

4.2 Security - design

Many systems today involve sensitive information, such as fingerprints, bank card pin numbers, and personal pictures. This is especially true for recognition systems, since the databases are sensitive due to the privacy consideration. If anyone accesses or steals it, it could be vulnerable to the privacy information and even the worse, if the database is been modified, it may create inaccuracy and confusion on the result. Therefore, encryption needs to be added to the system's database. This project records and stores three crucial pieces of personal information: the 2D video frame image, the 3D vertices point's data and the audio wave data. Since the data size varies between these three different data types, the range of data size goes from 5KB up to 20MB (for each sample). A suitable cryptography is important for efficiency, security, and process environment suitability. We studied three most well-known algorithms: Data Encryption Standard (DES), Triple Data Encryption Algorithm (3DES) and Advanced Encryption Standard (AES). It is known that DES, 3DES and AES all use XOR logic in the encryption system. However, DES is designed based on feistel networks, where each encryption process is relatively linear to the next layer. To extend that, with 3 unique DES connected together, 3DES is created. On the other hand, AES follows a totally different design principle, based on Subs-perm networks, with byte sub, shift row, mix column and Inv shift rows on each process [50].

The comparison of the three method is detailed in experiment section (Chapter 5). As shown in Figure 4.1, AES has the shortest encryption time, while 3-DES has the longest, suggesting that AES be the most suitable for this project. With this observation, we can summarize that Advanced Encryption Standard; possibly AES-128 or AES-192 should be selected for this

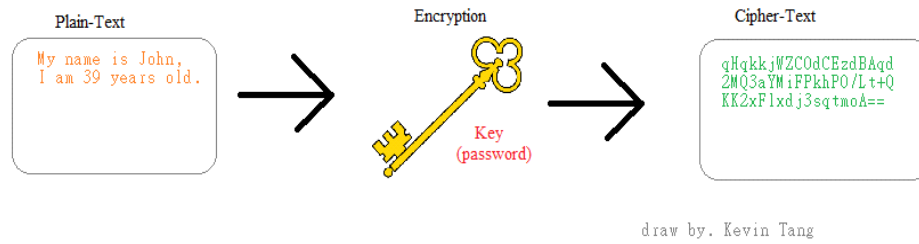


Figure 4.1. Cryptography.

project. The encryption process can be described as follows: a) normalization, b) packing, c) perform AES, and d) sending the decryption [51].

The 2D, 3D and audio data collected in the project contains personal information, include the facial image of the person, the audio print of person and 3D structure of a person’s facial structure, all this can be sensitive. To prevent a possible information leak or theft, the technique of “cryptography” becomes useful. This technique has existed for more than a few thousand years, since the age of humans fighting each other with stone and iron weapons. The idea of cryptography is to mainly focus on hiding a message, which is plain-text (shown in Figure 4.1). The sender will make the plain-text based on the encryption algorithm, and usually requires a Key to perform the Encryption. After this step, the plain-text becomes what we call cipher-text. In theory, cipher-text is not readable to anyone else besides the receiver who knows how to decrypt the cipher-text. This prevents the possible leak of sensitive information to other people who are not involved.

To prevent a possible information leak or theft, the technique of “cryptography” becomes useful. This technique has existed for more than a few thousand years, since the age of humans fighting each other with stone and iron weapons. The idea of cryptography is to mainly focus on

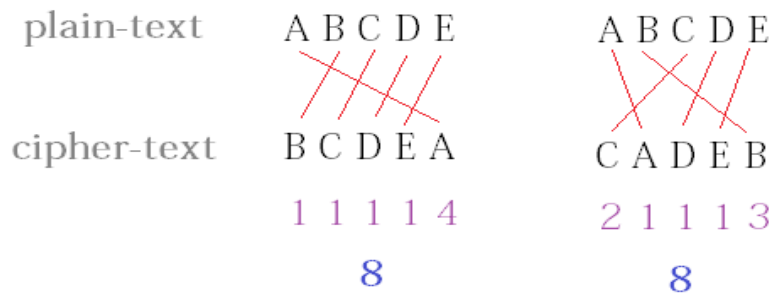


Figure 4.2. Shifting to create disorder.

hiding a message, which is plain-text (shown in Figure 4.1). The sender will make the plain-text based on the encryption algorithm, and usually requires a key to perform the encryption. After this step, the plain-text becomes what we call cipher-text. In theory, cipher-text is not readable to anyone else besides the receiver who knows how to decrypt the cipher-text. This prevents the possible leak of sensitive information to other people who are not involved.

The important functionality of encryption is disorder. In order to make a good encryption system, we want the maximum disorder. In Figure 4.2, we show a plain-text as ABCED. By moving a character from one position to the neighbouring position, or by moving it to a different, non-neighbouring position, we can see that even though both of them are the same number of shifts, the disorder levels are different. The left one can be easily changed back by shifting to the right once, but the right one won't be able to shift back by performing the same procedure to each character. This creates higher disorder in the right one.

In general, an Encryption System has four main functionalities as described below [19]

- I. *Secrecy* : Prevent anyone else besides sender / receiver to read the message.
- II. *Authenticity* : Recognition the message source, verify it is the correct user who sends the message.
- III. *Integrity* : Ensure that Cipher-Text has not been modified or changed by anyone else.
- IV. *Non-Repudiation* : Record the evidence that the user “did” send the cipher-text, and the receiver “did” received the cipher-text.

4.3 Overview of database

At this point, we can see many advantage of AES, but before we discuss the research on the weakness of AES, we should first look at how Hash Technology work, based on National Institute of Standards and Technology (NIST) announced in 1994, Message Authentication Code 5 (MD5) and Secure Hash Algorithm (SHA). It became the most commonly used non-invertible hash algorithms. In August 2004, Xiaoyun Wang (王小云) proposed a paper “*Collision Search Attacks on SHA1*” [17] regard the possible breaking method to MD5 and SHA-1, it can decreased the complexity of hash from 280 down to 269, which is 2000 times easier. The most important part is that MD5 and SHA-1 did not fully satisfy one of the key element of encryption, the *Non Collision*, which need to ensure that with different input won't have a common result, or as random as possible [36]. So far the research on AES shows AES did satisfy non-collision requirement, and we still haven't discover any attack which is able to break AES within a short period of time. However, many researches show we could possibility decreased the complexity of encryption by perform *Square Attack*, *Imp.Diff (Impossible differential) Attack*, *Partial Sums Attack*, and *Pushdown*

Attack [18], but its only decrease small level of security, plus the attack required to modify the chosen plaintext, which means it required the centre control to the user's computer. In this case, as long as the centre computer's owner did not release the password to the computer, the system theatrically is at a secure level. Finally, both DES, 3DES and AES have its own advantage and disadvantage as mentioned, to selective most suitable one required actually experiment, which will be shown in Chapter 5.

Chapter 5

Experiment

The experiment can be divided into three parts: security part, system suitability part and recognition part. Each part is been neatly performed and tested for multiple times until the conclusion could be drawn.

5.1 Experiment - Security

The difference between DES, 3DES and AES was compared. The 3D point's data was tested on the Windows platform to see the key generating time (KGT), encryption ratio (compared between the original file to the encrypted file), encryption time (dependent on the programming language used, which in our case was Visual C#), key length and estimate breaking time based on research. This leads to the conclusion of which encryption standard is the best suited for our study. Our input data was a 1.0 sec duration 3D vertices data text file with a size of 0.003759766 MB. The comparison between DES, 3DES and AES is shown in Table 5.1 and Figure 5.1. The result shows that AES achieves the shortest encryption time (average time for 10 tests) as 0.344 second compare with DES and 3DES.

Table 5.1. Encryption methods comparison on the data.

Algorithm	Number of test samples	Average size of test emotion data	Encrypted size	Key size	Average encrypting time	Average decrypting time
DES	10	3.76KB	3.76KB	56 bits	0.351ms	0.352ms
3DES	10	3.72KB	3.90KB	112 bits	0.372ms	0.380ms
AES	10	3.88KB	3.89KB	128 bits	0.344ms	0.347ms

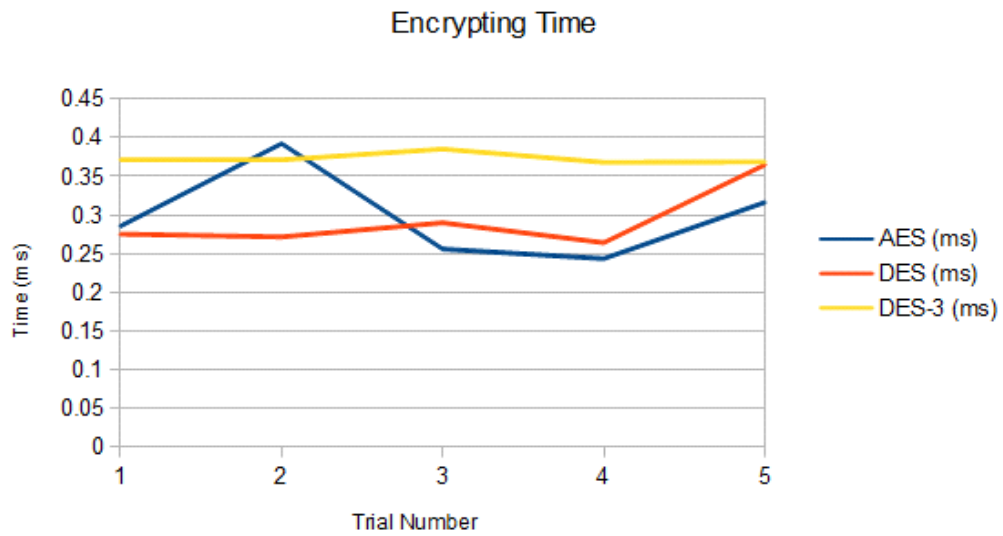


Figure 5.1. Result of encryption time.

The SSL transmission speed also needed to be tested, due to the fact that the data may be collected worldwide. To protect the security of the data, the encrypted data would be transmitted via SSL. In this case, we tested the SSL transmission speed on each set of data. Based on the comparison and research, we concluded that DES is the weakest encryption algorithm out of these three algorithms. Although 3DES reached a high level of security that is not breakable in a short period of time today, the encrypting duration is much longer compared to AES. Due to the fact that the 2D video image frame files can be up to several Megabytes, the overall database size could reach several Gigabytes. This makes the time crucially important. Considering the encrypting and decrypting time, AES is the most reasonable encryption algorithm choice for our study.

5.2 Recognition accuracy

The experiment we performed are involved with three different streams of data: 2D frames, 3D vertices and audio. Then, we perform the fusion of all three stream of data. Therefore, to perform the experiment, we start with 2D frames experiment, to perform the recognition accuracy experiment. We start with 2D visual frames. In each of the seven emotions (neutral, happy, sadness, Angry, fear, disgust and surprise), we randomly select 30 different individuals' frames out of 30 to 45 volunteers depending on emotions categories. The reason some emotions have fewer data collected is because some emotion is harder to preformed such as the emotion of fear. On the other hand, the emotion of happiness is the emotion that everyone is able to perform. The randomly selected 2D visual frames have been shown in Figure 5.2 for reference.

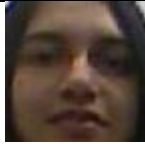

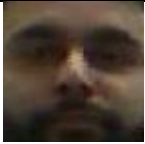
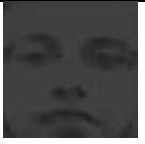




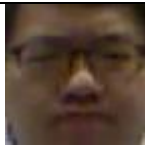

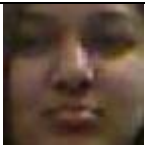

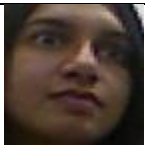




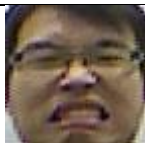



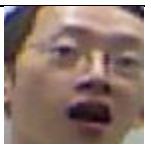

	Example1	Example 2	Example 30	Trained example
Neutral				
Happy				
Sadness				..
Angry				..
Fear				..
Disgust				..
Surprise				..

Figure 5.2 2D emotions frames.

As discuss in the previous chapter, we use PCA and then applied nearest neighbour classification, and the recognition accuracy is shown in the Table 5.2. Note that NU represents the emotion of Neutral, and HA: Happy, SA: Sadness, AN: Angry, FE: Fear, DI:

Table 5.2 Recognition accuracy vs. trained ratio.

Trained Ratio	NU	HA	SA	AN	FE	DI	SU
1	19.04762	20.63492	28.04233	27.51323	26.98413	28.04233	28.04233
2	30.95238	33.33333	32.14286	32.14286	32.7381	31.54762	29.16667
3	23.80952	36.73469	38.77551	38.09524	38.77551	38.09524	39.45578
4	31.74603	46.03175	45.2381	44.44444	44.44444	45.2381	45.2381
5	29.52381	47.61905	48.57143	51.42857	50.47619	50.47619	53.33333
6	32.14286	48.80952	46.42857	45.2381	48.80952	51.19048	53.57143
7	41.26984	49.20635	50.79365	50.79365	57.14286	58.73016	57.14286
8	35.71429	52.38095	57.14286	54.7619	54.7619	54.7619	59.52381

Disgust and SU: Surprise. When the trained ratio increased, the test sample ratio decreased, at the same time the recognition accuracy increase too. We can also view it in Figure 5.2, and it is clear to see the recognition accuracy increased when trained ratio is increased. Also, based on Figure 5.3, we can also see that neutral emotion has the lowest recognition accuracy among all seven emotions. This is possibly due to the fact the neutral emotion does not continued any specific feature that is different than others, which is more difficult to classify.

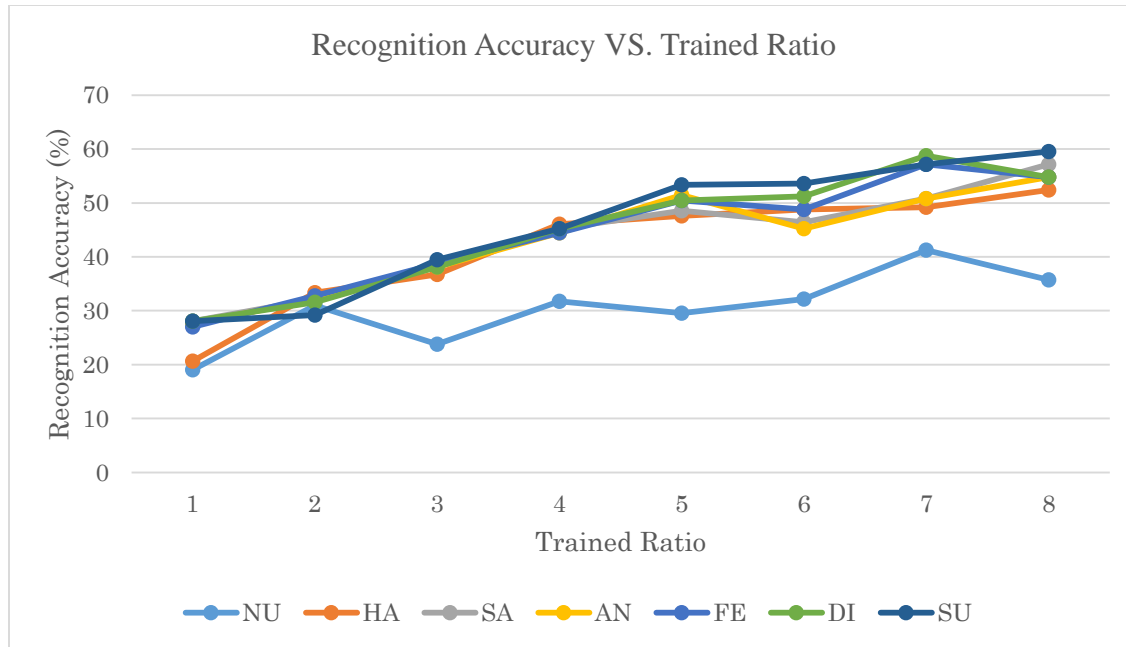


Figure 5.3 Recognition accuracy vs. trained ratio.

5.2.1 2D based recognition

Trained Ratio categorizes the recognition rates between degrees from 1 to 8, as shown in Table 5.3.

Table 5.3 Recognition accuracy (2D).

	Neutral	Happy	Sadness	Angry	Fear	Disgust	Surprise
Average	30.52%	41.84%	43.39%	43.05%	44.26%	44.76%	45.68%

The overall recognition accuracy across 20 experiments for PCA with nearest neighbor method using a ratio of 8 is shown in Table 5.4. The emotion of surprise achieved the highest recognition accuracy of 59.52%.

Table 5.4 Recognition accuracy (ratio = 8).

	Neutral	Happy	Sadness	Angry	Fear	Disgust	Surprise
Average	35.71%	52.38%	57.14%	54.76%	54.76%	54.76%	59.52%

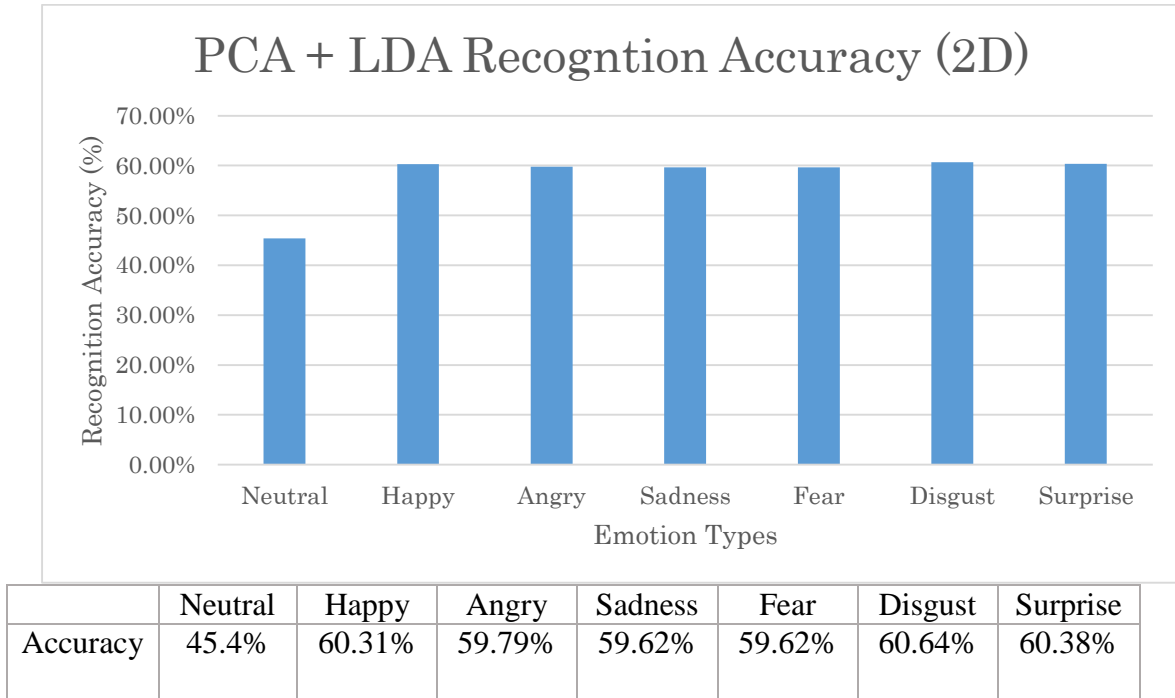


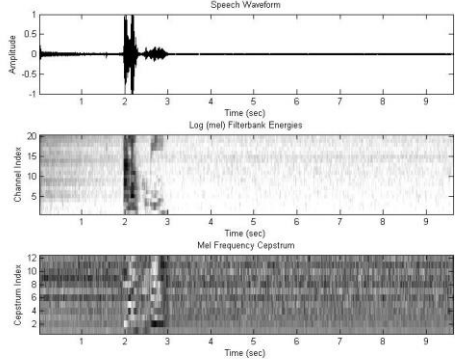
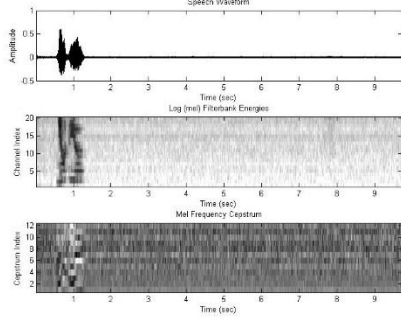
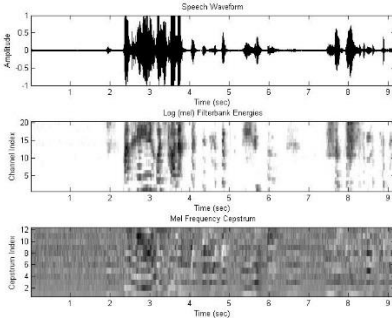
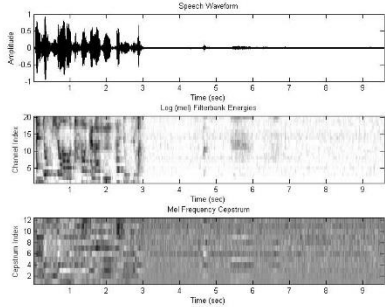
Figure 5.4 PCA + LDA recognition accuracy.

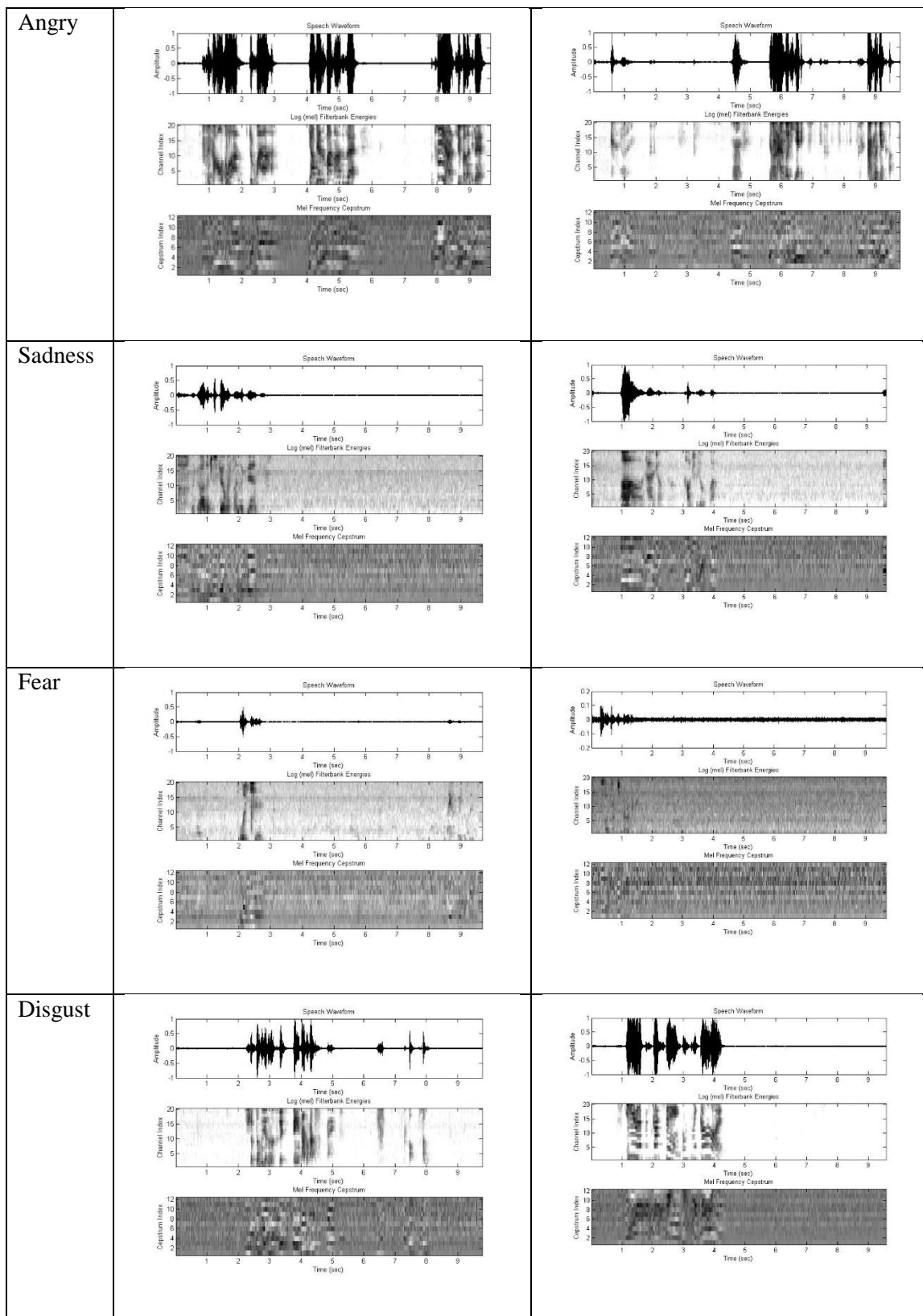
Next, we performed the recognition with PCA plus LDA, as shown in Figure 5.4. The recognition accuracy increased. Also, the emotion of neutral once again has the lowest recognition accuracy rate. The overall recognition accuracy by using PCA + LDA is 57.96%.

5.2.2 Audio based recognition

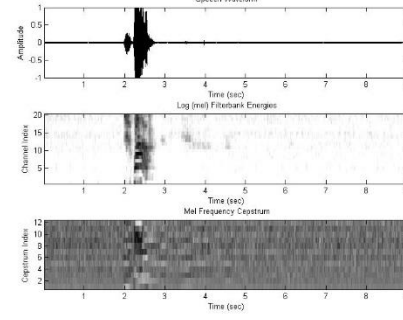
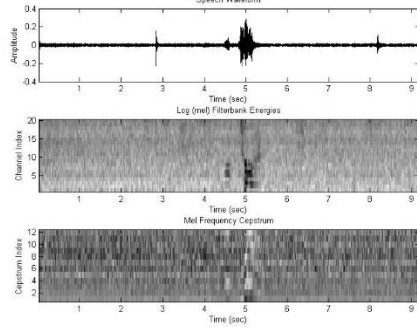
For the audio part, the experiment is based on MFCC analysis. The wave feature of time, log, and MFCC spectrum are shown below in Table 5.5. On the left column of the table is the trained model, and on the right column is the testing model.

Table 5.5 Audio feature in time, log and MFCC spectrum.

	Trained	Testing
Neutral		
Happy		



Surprise



Based on the 13 MFCC features and another 7 features per data, the audio recognition accuracy is shown in Table 5.6. The overall recognition accuracy is 60.88%. As we can see, our method achieves the highest accuracy for the emotion of angry amongst all seven emotions. The possible reason is that angry contains much higher energy in the frequency domain compared with the others.

Table 5.6 Recognition accuracy (audio).

Neutral	Happy	Angry	Sadness	Fear	Disgust	Surprise
56.80%	60.74%	66.97%	62.31%	59.11%	64.07%	62.18%

5.2.3 3D feature based recognition

Afterwards, we tested the 3D vertices part. In Table 5.7, we display all 121 feature points using the Open Scene Graph (OSG) environment. As we can see, the 3D modal shows significant differences for the emotion of surprise. For the 3D vertices recognition, we take all 121 points. LDA is not required since the dimensionality of our 3D data is relatively lower compared with the 2D frame data and audio data. The recognition accuracy is shown in Table 5.8. As we can see, the recognition accuracy is relatively lower than the 2D frame and audio parts. The overall recognition accuracy is 33.03%. The surprise emotion class achieved the highest amongst all seven emotions. This is possibly due to the significant difference at the mouth region which appears more widely open in surprise compared with other emotions.

Table 5.7 3D emotion modal example.

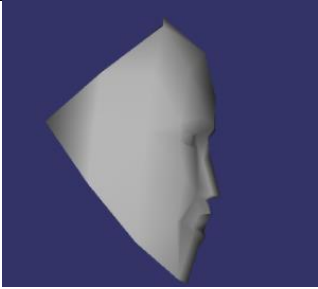
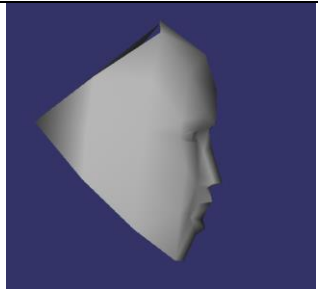

Happy	
Fear	
Surprise	

Table 5.8 Recognition accuracy (3D).

Neutral	Happy	Angry	Sadness	Fear	Disgust	Surprise
29.86	36.03	30.22	31.61	35.41	29.98	38.07

5.2.4 All three features based recognition (PMM-ER)

Finally we perform the fusion on all three categories of features. Based on the use of CCA as the algorithm and the proposed fusion algorithm as shown in Equation (3.20).

Table 5.9 – Recognition accuracy (2D + 3D + Audio).

Neutral	Happy	Angry	Sadness	Fear	Disgust	Surprise
75.08	82.17	83.42	81.37	78.15	81.94	82.91

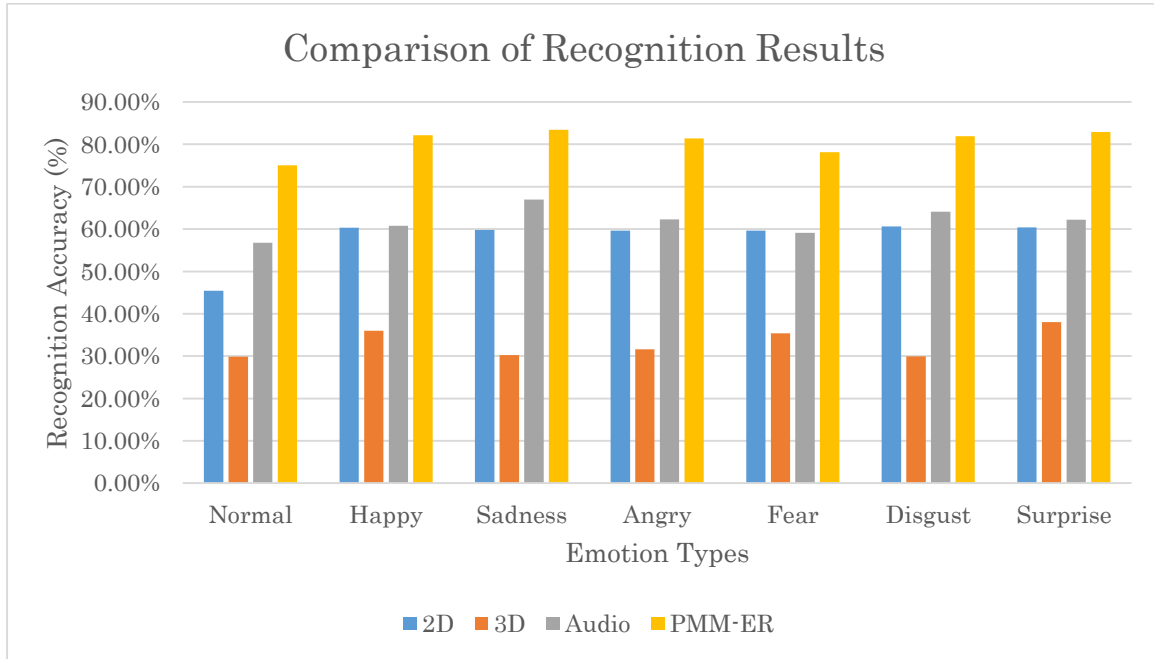
The recognition accuracy of proposed method is shown in Table 5.9. The highest recognition accuracy is the emotion of angry, it reached 83.42%. The lowest is the emotion of neutral, it only have outcome of 75.08%. The possible cause is because angry have relative high recognition accuracy in both 2D feature based and audio feature based recognition. On the other hand, the emotion of neuter is lowest in all three features, as the result it have the lowest recognition accuracy. We take a generally looking at other research from Korea and U.S.A's result [54][55], our proposed fusion result achieve 5% to 7% higher recognition accuracy result. (Shown in Table 5.10).

Table 5.10 Recognition accuracy for other research.

	Neutral	Happy (joy)	Angry	Sadness	Fear	Disgust	Surprise
E. Jang, B. Park and others [54]	81.5%	81%	N/A	75.1%	72.5%	N/A	84.6%
M. Suk and B. Prabhakaran [55]	N/A	67.8%	47.1%	50.0%	62.5%	69.2%	87.3%

5.3 Summary

Overall, we present the PMM-ER system based on its security protection features, the system suitability, and most importantly, the recognition accuracy. As summarized in Figure 5.5, by using multi-modal fusion the recognition accuracy increases compared with the approaches using the individual modalities. Overall, we achieve 80.72% recognition accuracy.



	<i>Neutral</i>	<i>Happy</i>	<i>Sadness</i>	<i>Angry</i>	<i>Fear</i>	<i>Disgust</i>	<i>Surprise</i>
<i>2D</i>	45.4%	60.31%	59.79%	59.62%	59.62%	60.64%	60.38%
<i>3D</i>	29.86%	36.03%	30.22%	31.61%	35.41%	29.98%	38.07%
<i>Audio</i>	56.8%	60.74%	66.97%	62.31%	59.11%	64.07%	62.18%
<i>PMM-ER</i>	75.08%	82.17%	83.42%	81.37%	78.15%	81.94%	82.91%

Figure 5.5 Comparison of recognition results.

Chapter 6

Conclusions

6.1 Summary

In this work, we present a protected multimodal emotion recognition system. The system is designed and tested on multiple criteria, including security features on the database. It has been tested on the commonly used encryption standards, DES, 3DES and AES. As a result, based on the encryption time and encryption speed we decide to use a combination of AES and DES to protect our database. Regarding the database, we perform pre-processing which includes normalizing the frame image, rotation on the 3D points, and noise removal on the audio data. This is done to standardize the data and make it easier to be trained. For the data training, we simply use a one-vs-all method. If ever the data size increases, we will train it again. We also turned the frame image into black and white to decrease the dimensionality. For the classification method, we employ principal component analysis (PCA), linear discriminant analysis (LDA) and canonical correlation analysis (CCA) as the three methods. The comparative study demonstrates that by using CCA in the fusion process and the combination between feature level fusion (FLF) with the decision level fusion (DLF) superior performance of our proposed framework. It is more suitable to this multimodal recognition system, because CCA has the ability to analyze higher dimensional problems in a shorter period of time, while returning a lower dimensional result. Overall, the result shows that the feature level fusion has approximately 81% recognition accuracy, which is 23% higher than the 2D frame image recognition and 48% higher than 3D vertices data. It shows our proposed fusion algorithm between multiple modalities helps increase the recognition accuracy.

To summarize our work, we have the following four major contributions:

- (1) We build a new multimodal database that contains real time 2D image frames, 3D feature vertices and audio emotion data. Our proposed multimodal database with extendable features allows it to be used in many different applications.
- (2) We provided a security feature on the database, based on the AES and DES encryption algorithms to protect the important personal information. The Security feature protects the original database and opens up possibilities of using this algorithm to improve personal data security.
- (3) We developed a combined pre-processing system on 2D images, 3D vertices and audio. (4) We proposed our own fusion method based on the combination of decision level fusion (DLF) and feature level fusion (FLF).

6.2 Future research

Overall, we investigate and develop a multimodal emotion recognition system. In this section, we will outline the future goals and potential enhancements for the work.

- Consider different modalities

As we discussed in the beginning, human emotion can be expressed from many different aspects, including face expression, voice, and heart rate. Considering more different aspects can result in more accuracy in reflecting the real emotion that a person is experiencing. In our project, we considered 2D facial image frames, 3D facial feature points and audio data. In future work and with the proper equipment, we can also consider heart rate, breath frequency and other aspects.

- Consider different emotion(s)

In our project, the seven emotions we used were happiness, sadness, anger, fear, disgust, surprise and normality. In reality, there can be many more emotions, such as frustration, anxiety, and annoyance. To increase the number of emotion types will require collecting the corresponding emotion samples at the database level. The methodology will be the same. The only difference is considering the different feature points and fusion algorithms based on the analysis of each additional emotion.

- New encryption standard

At the time when this paper was ready to be published, the headline news in the U.S.A showed that the National Security Agency (NSA) had the potential ability to break Advanced Encryption Standard (AES) by using quantum computing for analysis [52]. Based on prediction, it can decrease the time from 100 million years of analysis down to only 5 minutes. This drop in time comes as a warning for the current encryption standards. The possible research based on quantum computer encryption may lead to a choice of finding a stronger encryption method after AES.

- Consider a new device

In Oct 2013, Microsoft released a new gaming system, the XBOX ONE. It came with the new KINECT sensor, commonly known as KINECT 2 or KINECT for XBOX ONE.

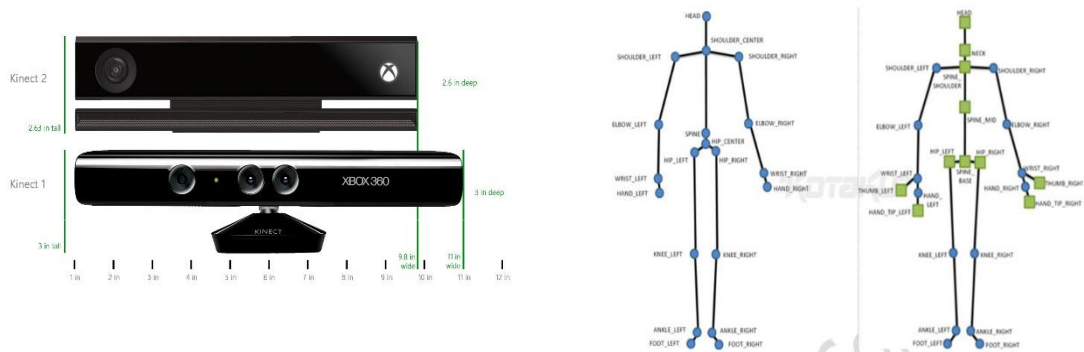


Figure 6.1. New KINECT model (KINECT for Xbox360 and XBoxONE).

KINECT 2 improves resolution on both 2D and 3D capture. It started using the Time of Flight technology (ToF), which increases the quality of an image.

References

- [1] S. Bentin, T. Allison, A. Puce, E. Perez, and G. McCarthy, "Electrophysiological studies of face perception in humans," *Journal of Cognitive MIT*, pp. 551-565, 1996.
- [2] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," 4 January 2012.
- [3] Y. Wang and L. Guan, "An investigation of speech-based human emotion recognition," *IEEE 6th Workshop on Multimedia Signal Processing*, 2004.
- [4] E. Bagherian, R. Wirza O.K. Rahmat, "Facial feature extraction for face recognition: a review," *Information Technology ITSIm, International Symposium*, Vol 2, pp. 1-9, 2008.
- [5] P. Ekman, "Strong evidence for universals in facial expression: a reply to Russell's mistaken critique," *Psychological Bulletin*, vol. 115, pp.268-287, 1994.
- [6] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Transaction. on Multimedia*, vol. 10, no. 5, pp. 936 - 946, August 2008.
- [7] W. D. Rencken and H. F. Durrant-Whyte, "A human-computer interface for a multi-sensor surveillance environment," *Advanced Robotics*, 1991. 'Robots in Unstructured Environments', 91 ICAR., Fifth International Conference on, vol.2, pp. 1761 - 1765, Jun 1991.
- [8] J. Fu, D. Miao, W. Yu, S. Wang, Y. Lu and S. Li "Kinect-like depth data compression," 2012 IEEE International Conference on Multimedia an Expo, 2012.
- [9] Y. Tie and L. Guan, "A deformable 3-D facial expression model for dynamic human emotional state recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 23, I, pp. 142-157, 2013.

- [10] Fredrik Gløckner. (2013, July). Networks [Online]. Available: <http://m43photo.blogspot.ca/2013/07/panasonic-lumix-dmc-3d1-review.html>
- [11] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in Signal Processing and its Applications CSPA 2011 IEEE 7th International Colloquium on. IEEE, pp. 410–415, 2011.
- [12] B. Chettr and K. B. Shah, "Nepali Text to Speech Synthesis System using ESNOLA Method of" Concatenation," International Journal of Computer Applications (0975 – 8887) Vol 62, No.2, 2013.
- [13] D. Biddulph, Biddulph and Balashek, "Automatic Recognition of Spoken Digits," Journal of the Acoustical Society of America Vol 24 No 6, November 1952.
- [14] S. Paulmann, M. D. Pell and S. A. Kotz, "How aging affects the recognition of emotional speech," Brain and Language 104, pp. 262–269, 2008.
- [15] B. Chettr and K. B. Shah, "Nepali Text to Speech Synthesis System using ESNOLA Method of Concatenation," International Journal of Computer Applications (0975 – 8887) Vol62 No.2, pp 24-28, January 2013.
- [16] S. J. Wang, C. H. Yang, "Computer and Network Security in Practice Applying to High-tech Society," 2011.
- [17] X. Wang, Y. Lisa, Y. H. Yu, "Collision Search Attacks on SHA1," Shandong University, China, February 13, 2005.
- [18] A. Sharifi, H. Soleimany and M. Aref, "9-round attack on AES-256 by a 6-round property," Electrical Engineering (ICEE), 2010 18th Iranian Conference on, vol., no., pp.226,230, 11-13 May 2010.
- [19] M. Gardner, "Codes, Ciphers, and Secret Writing," Dover Publications, Inc., 1972.

- [20] Microsoft (2012, Sep, 17). Face Tracking. [Online] Available : <http://msdn.microsoft.com/en-us/library/jj130970.aspx>
- [21] I. Patel and Y.S. Rao, "Speech Recognition Using Hidden Markov Model with MFCC-Subband Technique," Digital Object Identifier, pp. 168 – 172, 2010.
- [22] B. Guan; Y. He, "Optimal resource allocation for video streaming over cognitive radio networks," Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on, vol., no., pp.1,6, 17-19 Oct. 2011.
- [23] X. Nan, Y. He, and L. Guan, "Optimal resource allocation for multimedia cloud based on queuing model," in Proc. IEEE Int. Workshop Multimedia Signal Processing (MMSP'11), Hangzhou, China, Oct. 2011.
- [24] D. Fadi and A. Jorgen, "Fast and reliable active appearance model search 3D face tracking," in Proceedings of Mirage 2003, March 2003.
- [25] M. Paul and K. Fumio, "Milgram's Reality-Virtuality Continuum," IEICE Transactions on Information and Systems Vol.E77-D No.12 pp.1321-1329, 1994.
- [26] A. S. Mian, "Shade Face: Multiple image-based 3D face recognition," Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on , vol., no., pp.1833,1839, Sept. 27 2009-Oct. 4 2009.
- [27] G. Yin, D. Yang, Q. Wen, C Lai and J Shen, "Sincerity and User Avatar Research Based on Binocular Vision in Virtual Reality," Cybernetics and Intelligent Systems, 2006 IEEE Conference on , vol., no., pp.1,5, 7-9 June 2006.
- [28] 3dMD (1999). 3dMDface System. [Online] Available : <http://www.3dmd.com/3dMDface/>

- [29] NBC NEWS (Apr 8, 2013). Infrared camera takes 3-D images from miles away. [Online] Available: <http://www.nbcnews.com/technology/infrared-camera-takes-3-d-images-miles-away-1C9255119>
- [30] Y. Cui, S. Schuon.; S. Thrun, D. Stricker and C. Theobalt, "Algorithms for 3D Shape Scanning with a Depth Camera," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.35, no.5, pp.1039,1050, May 2013.
- [31] J. Fu, D. Miao, W. Yu, S. Wang, Y. Lu and S. Li, "Kinect-Like Depth Compression with 2D+T Prediction," Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on , vol., no., pp.599,604, 9-13 July 2012.
- [32] D. S. Bolme, J. Beveridge, M. Teixeira and B. Draper, "The CSU Face Identification Evaluation System: Its Purpose, Features and Structure," Computer Vision Systems, 2003.
- [33] N. Perlroth and D. Gelles, "Russian Hackers Amass Over a Billion Internet Passwords," The New York Times, Technology, Aug 5, 2014.
- [34] Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.31, no.1, pp.39,58, Jan. 2009.
- [35] D. Shark, (March 22, 2010), "HSL and HSV colour Map" [Online]. Available: http://en.wikipedia.org/wiki/HSL_and_HSV#mediaviewer/File:HSV_colour_solid_cylinder_alpha_lowgamma.png
- [36] H. Kourkchi, H. Tavakoli and M. Naderi, "An improvement of collision probability in biased birthday attack against A5/1 stream cipher," Wireless Conference (EW), p.p 444-448, 2010.
- [37] P. Ekman, Facial Action Coding System, 2002.

- [38] D. Pizarro and P. Bloom, "The intelligence of the moral emotions: A comment on Haidt," *Psychological Review*, Yale University, 110, 293-296, 2001.
- [39] O. Rudovic, "Coupled Gaussian processes for pose-invariant facial expression recognition," *IEEE Transaction*, 2013.
- [40] Y. Tong, J. Chen and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *IEEE Transaction. Pattern Anal.*, vol 32, no.2, pp.258-273. 2010.
- [41] Y. Li, S. Wang, Y. Zhao and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 7. Pp. 2559-2573, 2013.
- [42] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Transaction. Systems*. Vol 42, no.1, pp. 28-43, 2012.
- [43] C.H. Wu, W. L. Wei, J. C. Lin and W. Y. Lee, "Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion," *IEEE Transaction. On Multimedia*, DOI: 10.1109/TMM, 2013.
- [44] C. H. Wu, J. F. Yeh and Z. J. Chuang, "Emotion perception and recognition from speech," *Affective Information Processing*. New York, ch.6, pp.93-110, 2009.
- [45] D. Bitouk, R. Verma and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no 7-8, pp. 613-625, 2010.
- [46] A. Mehrabian and J.A. Russell "Evidence for a Three-Factor Theory of Emotions," *Journal of Research in Presonality* 11, pp 273-294, 1977.
- [47] M. Pantic (November 27, 2008). Imperial College London, "Facial Expression Recognition," [Online]. Available:

<http://ibug.doc.ic.ac.uk/media/uploads/documents/EncycBiometrics-Pantic-FacExpRec-PROOF.pdf>

- [48] Y. Kareem (December 1, 2011), School of Humanities and Sciences, Stanford University, “Analysis of 2,135 of the world’s known languages traces evolution of human communication,” [Online]. Available: <http://shc.stanford.edu/news/research/analysis-2135-world%E2%80%99s-known-languages-traces-evolution-human-communication>
- [49] S. Theodoridis, and K. Koutroubas, “Pattern Recognition (2nd Edition),” Elsevier Science (USA), 2003.
- [50] J. Lai, K. Liao, Y. Lai, and R. Chen “Design CAROM Module Used in AES Structure for Sub-Byte and Inv-Sub-Byte Transformation,” International Symposium on Biometrics and Security Technologies, pp.198-202, 2013.
- [51] M. Wang, C. Su and C. Horng “Single- and Multi-core Configurable AES Architectures for Flexible Security,” IEEE Transactions on Vol 18, pp. 541-552, 2009.
- [52] R. Broman (December 28, 2014). The Verge, “New documents reveal which encryption tools the NSA couldn’t crack” [Online]. Available: <http://www.theverge.com/2014/12/28/7458159/encryption-standards-the-nsa-cant-crack-pgp-tor-otr-snowden>
- [53] Microsoft Taiwan (2011) “10 分鐘 Kinect for Windows 應用入門” [Online]. Available: <https://msdn.microsoft.com/zh-tw/evalcenter/hh367958.aspx>
- [54] Eun-Hye Jang; Byoung-Jun Park; Sang-Hyeob Kim; Youngji Eum; Jin-Hun Sohn, "A Study on Analysis of Bio-Signals for Basic Emotions Classification: Recognition Using Machine Learning Algorithms," Information Science and Applications (ICISA), 2014 International Conference on , vol., no., pp.1,4, 6-9 May 2014

- [55] Myunghoon Suk; Prabhakaran, B., "Real-Time Facial Expression Recognition on Smartphones," Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on , vol., no., pp.1054,1059, 5-9 Jan. 2015

List of Publication

Conference Paper

- K. Tang, Y. Tie, C. Yang and L. Guan, “Multimodal emotion recognition (MER) system,”
IEEE Canadian Conference on Electrical and Computer Engineering, Toronto, Canada, May
2014