A Study of a Vision-Based Lip Movement Analysis for Hearing-Impaired Person

by

Muhamad Kamil Syahid Bin Talha

(1330610902)

A thesis submitted in fulfilment of the requirements for the degree of Master of Science (Mechatronic Engineering)

School of Mechatronic

UNIVERSITI MALAYSIA PERLIS

2014

ACKNOWLEDGEMENT

Alhamdulillah, praise to Allah, upon His blessings, I successfully completed my master study. First of all, I would like to thank to the Almighty Allah SWT for all His blessing. I would like to thank my project supervisor, Dr. Wan Khairunizam Wan Ahmad, for supporting me and giving advice on the project progress. He cared for the progress of the project in order that it would be completed on time with success. Opportunity has been given to me to explore by my own on making decisions based on practical thinking that eventually makes me take charge on my own project that will benefit me in the future project that I may undertake.

Other than that, I would like to thank my group members in working together to complete this project. Without their cooperation and patience this project will not be as a whole. Many problem occurred throughout the project are solve with the effort and cooperation contributed by all members of the group.

I also like to thanks my family, especially my father and my mother for their compassion and advices and give me strength throughout completing this project. Without their supports and motivations, I could never achieve all of this.

Last but not least, I would like to thanks to those who have contributed whether directly or indirectly in completing this project.

TABLE OF CONTENTS

CHAPTER 3 RESEARCH METHODOLOGY

CHAPTER 4 RESULT AND DISCUSSION

LIST OF TABLE

LIST OF FIGURES

Analisis Pergerakan Bibir Berasaskan Penglihatan Digital Untuk Kegunaan Orang Cacat Pendengaran

ABSTRAK

Kajian pergerakan bibir melalui pemprosesan imej ialah tafsiran pergerakan bibir manusia semasa bercakap. Pekak dan orang kurang pendengaran sering mempunyai masalah untuk memahami apa yang sedang diperkatakan di dalam perbualan. Bahasa isyarat mungkin berguna bagi mereka untuk berkomunikasi tetapi tidak semua orang boleh memahami bahasa isyarat . Untuk membuat komunikasi yang lebih menarik dan tidak mempunyai apa-apa halangan untuk berbincang dengan sesiapa sahaja , bacaan bibir adalah cara terbaik untuk berkomunikasi. Orang pekak dan kurang pendengaran boleh bercakap seperti orang yang normal , bagaimanapun, mereka tidak mendengar perkataan yang dituturkan oleh diri mereka sendiri. Sebaliknya , tidak ada sistem maklum balas yang kukuh untuk mereka. Ini adalah sebab mereka tidak boleh bercakap dengan sebutan yang betul. Untuk mengajar mereka menyebut perkataan dengan betul, fasih dan memahami orang lain dalam perbualan, sistem bacaan bibir yang boleh melatih mereka menyebut perkataan dengan betul perlu dibangunkan . Pengkaji terdahulu mencadangkan banyak kaedah untuk mengenali perkataan yang dituturkan berdasarkan pergerakan bibir. Salah satu kaedah adalah melampirkan penanda warna ke permukaan bibir dan penglihatan kamera digunakan untuk mengesan penanda bergerak. Penyelidik yang lain mengekstrak rantau bibir dengan menggunakan Aktif kontur Model (ACM) atau kaedah Snake untuk menarik keypoint di sekeliling tepi bibir. Dalam kajian ini, satu sistem untuk mengesan dan mengenal pasti perkataan yang dituturkan berdasarkan pergerakan bibir telah dicadang. Kamera digunakan sebagai sensor penglihatan dan teknik pemprosesan imej digunakan untuk mengekstrak dan menjejaki bibir. Jarak mendatar bibir dan menegak bibir digunakan untuk mengukur elips manakala sistem mengesan pergerakan bibir. Trajektori elips adalah disampel semula untuk 10 mata dipanggil titik ciri-ciri. Pangkalan data mengandungi 10 perkataan yang kerap disebut di hospital telah direka berdasarkan pengagihan 10 mata ciri. Dua jenis pangkalan data perkataan telah direka iaitu pangkalan data perkataan individu dan pangkalan data perkataan universal. Pangkalan data perkataan individu ditakrifkan sebagai titik ciri bagi setiap orang yg diuji dengan bercakap perkataan berulang kali. Sementara itu, pangkalan data perkataan universal ditakrifkan sebagai titik ciri bagi semua perkataan tidak kira sesiapa sahaja orang yang diuji. Eksperimen mengenal pasti perkaataan dituturkan telah dijalankan dan sistem ini mengenal pasti perkataan yang tidak diketahui dengan kadar ketepatan sebanyak 92.47% dengan menggunakan pangkalan data individu. Sementara itu, sistem ini mengenal pasti perkataan yang tidak diketahui dengan kadar ketepatan sebanyak 90.39% dengan menggunakan pangkalan data universal.

A Study of a Vision-based Lip Movement Analysis for Hearing-impaired Person

ABSTRACT

The study of vision based lip movement is the interpretation of human lip movement while speaking. Deaf and hard of hearing people often have a problem being able to understand what being talking in the conversation. Sign language may be useful for them to communicate but not everyone may understand the sign language. To make the communication more interesting and does not have any obstacle to talk to anybody, the lip reading is the best way to communicate. Deaf and hard hearing person can talk like the normal person, however, they do not hear the word spoken by themselves. On the other hand, there is no sound feedback system for them. This is a reason they could not speak with the correct pronunciations. To teach them pronounce the word correctly, fluently and to understand others in the conversation, the lip reading systems that could train them pronounce the word correctly needs to be developed. Previous researcher proposed many methods to recognize the spoken word based on the lip movement. One of the method is attach the colour marker onto the lip surface and a vision sensor is used to track the moving marker. Other researchers extract the lip region by using Active contour Model (ACM) or a snake method to draw the keypoint around lip edges. In this studies, a system to track and recognize the spoken word based on lip movements is proposed. A camera is used as a vision sensor and an image processing technique is employed to extract and track the lip. The lip horizontal and vertical distances of the lip are used to measure the ellipse while the systems track the lip movements. The trajectories of the ellipse are resampled to 10 point called the features point. The word database contains 10 spoken word frequently speak at the hospital are designed based on the distribution of the 10 feature points. Two types of word database have been designed which are the individual word database and the universal word database. The individual word database is defined as the distributed feature points of each subject by speaks the words with repetition. Meanwhile, the universal word database is defined as the distributed feature point of all word regardless who is the subjects. The recognition experiments are conducted and the system recognizes the unknown words with the recognition rate 92.47% accuracy by using the individual database. Meanwhile, the system recognizes the unknown words with the recognition rate 90.39% accuracy by using the universal database.

CHAPTER 1

INTRODUCTION

1.1 Project Overview

The World Health Organization (WHO) has estimated that 278 million people worldwide have moderate to profound hearing loss in both ears (Paulraj M P et al, 2008). The rising number of hearing impairment person worldwide with all level of age is due to the growing global population and longer life expectancies. The deaf people have developed their own cultures and communicating method among themselves as well as with ordinary person by using sign gestures. Sign gestures are different from the spoken language. It is a non-verbal visual language but serving the same function for communication. Recently research works on lip gestures have gained a lot of attention by many researchers in the area of the computer vision, the natural language processing and the pattern recognition (Mehrotra et al., 2009; Saitoh & Konishi, 2010; Lay et al., 2012). Lip gesture systems require the knowledge of shapes, motions, orientations and facial expressions (Tian et al., 2000).

Hard of Hearing and deaf people always have a problem to lip-read and understand other people. Often hard of hearing and deaf people feel they are actually ignored by other people or left out of conversation. This is very difficult for hard of hearing and deaf people to deal with, they could feel embarrassed to ask people what they are talking about and to ask for things to be repeated again. There are variety of ways hard of hearing person can gain access to communication and information. Communication support includes both human aid and technical. Human aid include interpreters, note takers and lip-readers. Interpreter who translates the sign language

must be qualified. Lip-readers help to train hard of hearing people to learn how to lip-read other person. In this communication aid, the interpreter who is lip-reader will sit face to face with the hard hearing person. The interpreter will correct the lip movement of the hard hearing person while speaking a word.

Speech recognition is not necessary only auditory. When a listener can see the speaker, visual information from lip movement is used in the process of speech recognition. The speech recognition based on the visual information was first introduced by McGurck and MacDonald (McGurck and MacDonalad, 1976). Development of command system based on speech to interface user with the computer is very useful to let the user control and interact with the computer naturally. This system development will provide more flexibility to the user to interface with the computer compared to the conventional interface by using mouse and keyboard. However, most of these systems are based on audio signals and are sensitive to the ambient noise, the signal strength and the acoustic conditions. To overcome this limitation, speech data that are orthogonal to the audio signal such as visual speech information can be used. The systems that combine the audio and visual modalities to identify utterance are known as audio-visual speech recognition (AVSR) system (Ayaz A. Shaikh, 2010). Visual speech recognition (VSR) system refers to the system which utilizes the visual information of the movement of the speech articulator such as the lips, the teeth and somehow the tongue of the speaker. The advantages are that such a system is not sensitive to the change in acoustic conditions and ambient noise, does not require the user to make a sound and provides the user with a natural feel of speech and dexterity of the mouth.

This project proposed to develop a lip reading system based on the real-time video streaming. From the video data, the point of interest of the subject's lip is tracked consecutively. Resampling method is introduced to analyse lip movements data and followed by the development of the spoken word database which focuses on the conversation in hospitals. The overall goal of this research is to enable deaf and hard of hearing person to practice speaking the correct pronunciation of the word spoken.

1.2 Motivation

Speaking, chatting or lip movement producing sound that can be understood is called communication. A normal people will communicate with each other by using the language that they understand. In this research project, Malay and English languages have been used in the experimental studies.

The purpose of using Malay and English languages is because of Malaysian mostly using these two languages in their daily communication. As a normal person it's not difficult for us to communicate, but for the deaf and the hard hearing person is difficult for them when getting treatment at the hospital to communicate with the doctor and the nurse.

To ensure the welfare for deaf and hard hearing person is taken care of, the communication between them with us needs to be improved. Currently, the sign language might be the best communication tool to communicate with them. But, not all of us know how to use the sign language. Besides that, for the eldest to learn the sign language have difficulty in memorizing it.

Lip reading technique had been use for century to recognize what a person is trying to speak. Some of the deaf and hard hearing persons practice lip reading to understand conversation. But not everybody can predict the spoken word correctly through the lip reading.

Lip reading had been practice by deaf and hard hearing people by guided by the interpreter. Currently they need to be face to face with the interpreter and the interpreter will correct their lip movement. This process of learning has some disadvantages because the deaf person requires a trainer to train them and it will feel uncomfortable by facing each other nearer to see lip moving.

1.3 Problem statement

The main problem in this research is to develop a marker-less technique for the detection and tracking of the lip movements. Before tracking the lip movements, the most difficult part is the extraction of the lip region because the lip does not have much colour contrast with the skin colour. Detecting lip region is a complex process because of the high variability range of lip shapes and colour (Skodras, et. al., 2011)

To recognize the lip movement, data from training lip movement are stored in database. A method to organize and recognize the group of data for respective word spoken is needed. The application of matching criteria allows comparison to be made between feature representing unknown data and feature within the database.

This is due to the fact that current technique is manual training by the trainer facing the deaf person to teach them how to speak (Auerbach, Jill, 1984). This type of training is so tiring because in one session student needs to really focus on trainer lip to follow lip moving based on the word spoken. If the virtual lips apply, a person can learn and train lip reading anytime and anywhere.

1.4 Objective

The objective of this research are described below:

- To investigate image processing techniques for detecting lip movement.
    - Lip is detected by using marker less technique. Colour of the lip surface's area has to be defined to extract the lip area. Lip area will then be smoothen it shape by applying several type of morphological process to acquire the best shape of the lip.
- To analyze data analysis by resampling the raw data of the lip movement
    - Raw data that is collected from several subject for 10 word will not have same number of frame while recording lip movement. This is due to the speed of the subject speaks the word and the size of the lip. To overcome this problem, data will be resampled to extract movement feature from the raw data and standardise the number of frame for all testing data.
- To recognize words by using movements of the lip.
    - Input data from the unknown spoken word will be checked for it similarity with the stored words in the database. The proposed similarity measure approach is the Euclidean distance.

1.5 Scope

The main scope for this research is to detect and track the lip movement for the respective word spoken. To achieve this goal, the lip region need to be extracted and the threshold value for all subjects need to be determined and differentiated. After the lip region has been recognized, it will be smoothen by the morphological process. The key point will be developed on the lip edges. This key point will be used for tracking the lip movement when speaking. Lip movement data will be resampled to extract the lip movement feature. In the experiments, the selected spoken words are the word that are commonly in the hospital when the patient communicates with the doctor.

1.6 Project Expectation

The main target of this project is to develop a system that may help deaf and hard hearing person to learn effective way to speak. This system uses the vision sensor to track the real time image of the lip while speaking the word. There are no colour marker attached to the lip of the subject. The proposed system will automatically track the lip while the subject speak and recognizes the word spoken.

1.7 Report organization

This thesis comprises of five chapters and is organized following the order of work implemented. A brief overview and background of the current works, including the problem statements, objectives, and scopes of the thesis are described in this chapter.

An intensive literature review on the research of the lip reading is presented in chapter 2. A definition of the lip reading and its characteristics are discussed. The chapter also elaborates the trend of research on the extraction of the lip features and analysis methods from the past to the current.

Chapter 3 explains the methodologies used in this research. It start with the detailed description of the proposed lip tracking systems. Then a full description on the proposed vision tracking and recognition of the lip movement is explained. Some preliminary results are also presented in this chapter to demonstrate and prove the effectiveness of the proposed algorithms.

Several experiments are conducted in Chapter 4 to validate the robustness of the lip tracking systems for the word used in the hospital.

Chapter 5 concludes this thesis with the summary and the recommendations for future works.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Many research works has been developed in the field of lip reading. Word recognition based on the vision system has become mainstream recently due to the research and development of the vision sensor. In the human language communication process, lip shape changes and facial expression of the teller contains rich language information (Lay et al., 2012). In addition, some researchers have developed speech recognition based on the shape of the lip and also distance chance between the upper lip and lower lip. To obtain geometric feature of the lips, lip contour, several curves and special point on the lip were collected to develop lip template that approximately actual lip shape (He, 2010). Other researchers recognize movement and distance between upper lip and lower lip to indicated condition of mouth-opened or mouth-closed (Chen et al., 2004).

In another research, (Skodras and Fakotakis, 2011) demonstrates automatic and accurate lip segmentation algorithm. Automatic k-mean were used to segment lip pixel based on the colour information of the lip. After morphological process and ellipse fitting, mouth corner need to fine tuning by using Harris corner detector. Corner detector makes the lip detection precise for lip reading. While, (Talea and Yaghmaie, 2011) presented lip detection from combination of two colour extraction method. Author proposes the combination of red exclusion algorithm and colour transform by using red and green colour information. This two combinations extraction method used

to show the exact shape of the lip while in the room condition with the disturbance of light and shadow below lower lip.

This chapter begins with the terminology of deaf and hard hearing persons. The type of phonemes and visemes in lip reading will be explained.

2.2 Lip reading

Output from the experiment indicates that lip reading can be applied efficiently in constrained-lexicon verbalization apperception (J.C Wojdel and L.J.M Rothkrantz, 2001; S.Lucey et al., 2001), apperception of verbalization uttered by verbalization impaired (G. Potamianos and C. Neti, 2001) and additionally in case of perpetual verbalization signal (P. Wiggers et al., 2002). Methods and techniques developed in developing automatic lip-reading useable in the computer generated facial animation and multimodal verbalization synthesis (T. Chen, 2001). The lip forms of movement and other visually distinguishable transmutation in articulatory system are represented by different researchers in multitude of possible geometric and non-geometric models (G. Potamianos et al, 2003).

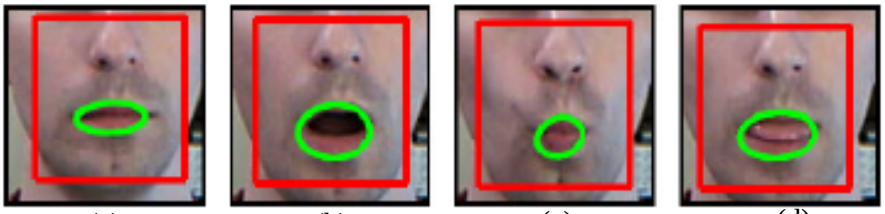
According to Conrad (1979), the capacity for lip reading seems to be resolute by the person's degree of hearing perceiving and the caliber of perspicacity and of verbalizing. However, the studies that have fixated on establishing the cognation between the degree of auricular discerning and the caliber of lip reading have not reached unanimous conclusions. Some have observed that hearing people are customarily more preponderant lip readers than hearing impaired people and therefore they have concluded that the more the loss of aurally perceiving the more difficult lip reading will be.

There are 43 type of phonemes in English language, while only 28 different mouth shape differentiate them (A. Greenwald et al, 1984). Alphabet‘d’ and‘t’ while ‘f’ and ‘v’ both share the same mouth shape. Therefore, the art of lip reading for humans is context sensitive. To recognize word spoken, mouth shape only cannot be indicate the word spoken by the speaker, but it needs the knowledge about the environment and surrounding of the word spoken.

2.3 Related research work

Recently, several researchers have focused on developing lip reading system in recognising words spoken. There are several methods have been developed to determine the lip region and classifier to classified word spoken based on the lip movement. The precision of human lip readers can be ameliorated significantly when constrained erudition of the language is introduced and that in less than ideal condition, good human lip readers are able to exploit language well and amend their precision (Lan et al, 2012).

Dalka and Czyzewski (2009) introduced lip and gesture recognition for interfacing multimodal human-computer. Author developed a system that sanctions utilizer work on a computer by utilizing mouth gesture and movement. This system was called LipMouse. This system helps disabled and paralyzed person to control mouse cursor by lip movements. Face recognition is predicated on a cascade of boosted classifier utilizing Haar-like feature while lip been identified on the lower region part of the image. Artificial neural network is utilized to recognize lip gesture. Besides, precise lip region is identified by the segmentation of the lip region by utilizing fuzzy clustering. Figure 2.1 show the sample result of mouth region and lip shape detection.



(a) (b) (c) (d)

Figure 2.1: Sample result of mouth region (red rectangle) and lip shape detection (green ellipse) for the lack of the gesture (a) and for three gestures recognized: opening the mouth (b), making the mouth into an “o” shape (c) and sticking out the tongue (d) (Dalka and Czyzewski, 2009).

He (2010) used a lip template to develop a method to extract characteristic parameters of lip. Prediction of lips shape on the criteria of the least square was determined using a dynamic clustering algorithm and ameliorated ant colony algorithm obtained the dynamic state sequence that described the lip movement. Dynamic Time Warping (DTW) algorithm performs the recognition of the lip reading dynamic pattern. Chinese syllable a, u, e and i for the image of lip shape sequence have been accumulated. Six different lip shapes are divided which are close, open slightly, open half, open fully, grin and pout. The reference template were trained by the former 20 sets of data for each syllable. The latter 20 sets are habituated to be apperceived. Recognition result are that recognition precision of syllable a is 90%, syllable u is 75%, syllable e is 65% and syllable i is 70%. Figure 2.2 show the feature image of sequence for syllable *a, u, e* and *i.*