Lip Reading to Text

Mogammat Waleed Deaney

Project presented in partial fulfillment
of the requirements for the degree of
Bachelor of Science (Honours)
at the University of the Western Cape

Supervisor: Prof.I.M Venter
Co-supervisor: Mr.M Ghaziasgar
Mentor: Kenzo Abrahams
Co-mentor: Nathan de la Cruz

28 March 2014

Declaration

I, Mogammat Waleed Deaney, declare that this project "*Lip Reading to Text*" is my own work, that it has not been submitted before for any degree or assessment at any other university, and that all the sources I have used or quoted have been indicated and acknowledged by means of complete references.

Signature: .... Date: ....

Mogammat Waleed Deaney.

Abstract

Lipreading also known as speechreading, is a visual way of “listening” to someone. This is done by looking at the speakers face to follow their speech patterns in order to recognise what is being said. This project will attempt to develop a computerized lip-reader. The system will observe the user’s lip-movement through a sequence of frames and convert the recognised sound/word to text which will be displayed on a screen. The project will regard the mouth as the region of interest for lip reading. All other features such as movement of the tongue, hand and facial gestures is considered to be beyond the scope of the project. The subject will be asked to articulate certain sounds or letters which need to be correctly identified by the system.

Key words

Visual speech recognition

Lipreading

Speechreading

Acknowledgments

I would like to thank my supervisors Prof.I.M. Venter and Mr.M.Ghaziasgar for their support and encouragement during my Honours year. Without our weekly meetings, this work would not have been possible. Without their help I would certainly not be where I am today.

Contents

List of Tables

List of Figures

Glossary

LBP	Local Binary Pattern
LBPH	Local Binary Pattern Histogram
SVM	Support Vector Machine

Chapter 1

Background

Speech recognition is not purely auditory. A speaker produces visual information which can be used by the listener to interpret what the speaker is saying. This is considered to be part of the speech recognition process. The first reports of visual information being used for speech recognition were contributed by McGurck and MacDonald [McGurk and MacDonald, 1976] [Mark Barnard and Owens, 2010].

This project is an attempt to recognise lip movement as speech and display the result as text. It entails developing an application that uses a web camera, which will be pointed at a person's face, to lip read that person. This will involve recognizing simple vowels or sounds at first. It can later be extended to words and sentences, time permitting. The application will then take the recognized vowel, word or sentence and convert it to text to be displayed on the screen.

Chapter 2

User Requirements

2.1 Introduction

This section describes the problem from the user's point of view. It states the problem domain and the functionality of the program.

2.2 Users view of problem

Lip reading is the process of understanding speech by visually interpreting movement of the mouth. The user needs a system that will computerize the process of lipreading.

2.3 Description of the problem

A system is required that can automatically lip read letters, sounds or words without input from the user. The objective is to create an automated lip reader that will determine what the subject has said and display it on the screen.

Speech recognition translates spoken words into text. It is difficult for the software to identify a particular speaker when another person is speaking at the same time. Another disadvantage is that it does not work well in noisy environments. The sound of particular words can sometimes be similar for e.g ”one” and ”won”. The words can be mistaken for one another as they sound the same but have distinctly different meaning and lip movements. The “lipreader” will address the disadvantages of speech recognition.

2.4 Expectations of the software solution

The software should be able to detect the subjects mouth-movements using a web camera and be able to recognise simple sounds or letters articulated by the subject. The program should display the recognised sound or letter as text to the user. It should also allow the user to stop and start the program at any given time.

2.5 Not expected of the software solution

The software will not track multiple subjects and is not expected to monitor subjects who are not directly facing the camera. The subject being lip-read is required to be as still as possible. The camera should have a clear uninhibited view of the user's face. The lip-reader application will not do real-time lip reading.

2.6 Conclusion

The above section describes how the user requires a system that will allow them to recognise speech visually. It also describes the expectations of the system to define the scope of the project. Chapter 3 provides the designers interpretation of the problem in order to help identify the most beneficial means for implementation.

Chapter 3

Requirements Analysis Design

3.1 Introduction

In the previous chapter, the user specified the need for a visual speech recognition system. This chapter will explain how the designer will deal with the user's requirements. This is done by breaking down the problem into high level constituent parts and identifying all the relevant details.

3.2 Interpretation and breakdown of the problem

The input of the application will be a live video stream of the subject speaking. The subject will have to face the camera directly and not move his/her head around while the program is running. When the subject speaks he/she will have to articulate well. The webcam will capture the change of mouth movements when the subject speaks.

In order to accurately identify the change in mouth movements, preprocessing of the the video snippet will have to be done. Some of the preprocessing will be done with OpenCV as its implementations are continuously being improved and or added upon. OpenCV is an open source computer vision and machine learning software library which allows for a simple-to-use computer vision infrastructure and has over 500 different functions available [Bradski and Kaehler, 2008].

3.3 Complete analysis of the problem

3.3.1 Face and Mouth Detection

The application will accept video input from the webcam frame by frame. For this project the mouth is considered as the region of interest. In order to detect the mouth, the subjects face needs to be located. This is done by implementing

OpenCV's face detection library. It uses the Haar-classifier which is a tree-based technique used to detect rigid objects [Bradski and Kaehler, 2008]. The core basis for Haar-classifier object detection is the Haar-like features [Wilson and Fernandez, 2006]. Haar-like features are a contrast in values across a rectangular area of pixels showing both light and dark areas. Both the face and mouth have Haar-like features and therefore can be detected using this algorithm and is in Figure 3.1.

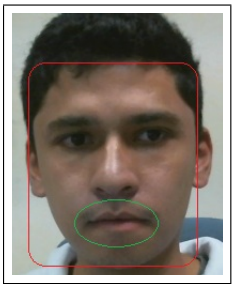

Figure 3.1: Locating Face and Nose

3.3.2 Pre-processing Frames

Once the mouth is detected it is segmented from the rest of the frame. The image of the mouth is then greyscaled in preparation for our next operation. Next the Local Binary Patten (LBP) operator will be applied on the image. LBP as an image operator is used to extract LBPH (LBP histogram) features for texture description[Yang and Ai, 2010]. This will show the valleys and ridges of the image and could be useful in extracting the features we need from the mouth.

3.3.3 Categorizing Frames

A Support Vector Machine (SVM) will be used to categorize the frames. The SVM will be fed images to train and test the system. The frames will be classiefied as either a neutral where no movement occurs or as a mouth movement which can be matched to speech. A basic SVM takes a set of input data and predicts for each given input which of the two possible classes forms the output, making it a non-probabilistic binary linear classifier. The decision is made solving a linearly constrained quadratic programming problem [Osuna et al., 1997].

3.3.4 Displaying the Speech as Text

The system will allow the user to see what the subject has said on the screen. The result will be displayed as soon as the user finishes the sound/letter.

3.4 Other Similar system

Automated lip reading is not a precise art as people speak differently. Lip-reading also becomes a subconscious process for someone who is able to speak and listen clearly over a period of time.

Harshit Mehrotra, Gaurav Agrawal and M.C. Srivastava created a lip reading system. In their paper they proposed a solution based on automatic lip contour tracking. Their aim was to recognize characters of the English language after a lip contour is accurately detected on each frame. This method worked well on a limited character database. [Mehrotra et al., 2009]

3.5 Suggested solution

The solution will detect mouth movement and should be able to recognise simple letters. The solution could later be extended to capture words and sentences leading to a real-time lip reader. The tools needed for the implementation are a web camera and a computer with OpenCV installed.

3.6 Conclusion

This sections states how the designer interpreted the problem. A generic framework for a solution was given in order to solve how the “lip reader” would proceed in visually recogising speech. The next chapter will describe how the results will be displayed to the user.

Bibliography

Bradski, G. and Kaehler, A. (2008). *Learning OpenCV.*

Mark Barnard, E.-J. H. and Owens, R. (2010). Lip tracking using pattern matching snakes.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices.

Mehrotra, H., Agrawal, G., and Srivastava, M. (2009). Automatic lip contour tracking and visual character recognition for computerized lip reading.

Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: an application to face detection. In *Unknown*, pages 130–136.

Wilson, P. I. and Fernandez, D. J. (2006). Facial feature detection using haar classifiers. *Journal of Computing Sciences in Colleges*, pages 127–133.

Yang, Z. and Ai, H. (2010). Demographic classification with local binary patterns.