EECS442 F14 Project Proposal
Visual Speech Recognition

Tian Zhang | tzha | CE | Senior
Yi-Sheng Hsieh | yshsieh | EE:s | grad school 2nd year
Tongshuang WU | twuac | CE | Junior

I. INTRODUCTION

Topic: Visual Speech Recognition

Visual Speech Recognition is a novel application in computer vision, combining the usage of image processing, artificial intelligence, object detection, pattern recognition, statistical modelling, etc. Using both visual information as an additional source for audio extraction, it has great impacts in occasions where the audio information cannot be effectively conveyed.

Several studies have touched this topic. However, most of the current audio-visual speech recognition technology focuses on using visual clues as aids to audio-visual automatic speech recognition (AV-ASR) system, which still rely heavily on audio, and the advantages of visual detection hasn't been fully taken. We therefore propose to build a visual speech recognition (VSR) system, which uses visual information (e.g. lip motion) as the stand alone source to recognize/classify words being spoken, and audio information as an optional auxiliary method for accuracy improvement. We will start from simple digits recognition. When its accuracy is confirmed, we will continue to try some temporarily discussed advanced features. We believe it can potentially be further developed for more social-related usage, such as aids for deaf and hard-hearing users in terms of lip-reading.

Basic Feature :

Recognize digits from lips motion

Advanced Feature:

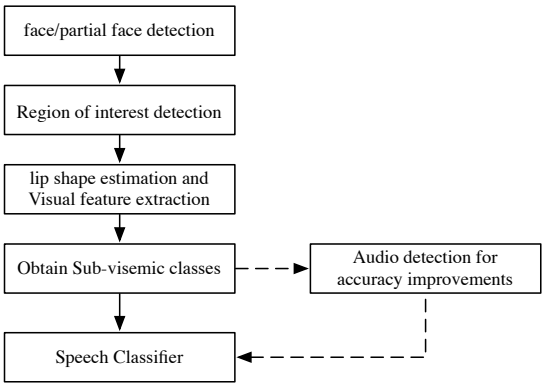
1. Combine with audio detection to analyze if the accuracy can be further improved
2. More complex and meaningful sentence, e.g. daily greeting, useful command (turn on, off, etc.)

Scenario of application

1. Deaf / auditorily handicapped aids
2. Skype in quiet place, like library.
3. shout out command (using Siri or controlling a car) in a noisy environment, like dialing in car.

II. TECHNICAL PART

We propose to process the visual language detection mainly with the lip motion recognition, which is a branch of partial face detection. Some related concepts and works are listed in sequential orders below. We also consider audio frequency spectrum classification as an auxiliary method for performance improvements.



1. Face/partial face detection

Face detection is a computer technology that determines the locations and sizes of human faces in digital images. It detects face and ignores anything else, such as buildings, trees and bodies. Face detection can be regarded as a more general case of face localization. In face localization, the task is to find the locations and sizes of a known number of faces (usually one). In face detection, face is processed and matched bitwise with the underlying face image in the database.

2. Region of Interest (ROI) detection

A region of interest (often abbreviated ROI), is a selected subset of samples within a dataset identified for a particular purpose. We want to use unsupervised approach to the detection of regions of interests from a image. We define the regions of interest as highly probable rectangular regions including mouth and lips. Two major lip localization/detection technique are Model-based lip detection and Image-based lip detection. Model-based lip detection uses template and models such as Active Shape Model (ASM) and Active Appearance Model (AAM) to detect lip location while Image-based lip detection make the use of spatial information, pixel color….etc.

3. Lip shape estimation and Visual feature extraction (PCA, LDA, DCT, DWT)

It focuses on the automatic extraction of Speech lip features from natural lips. The method is based on the direct prediction of these features from predictors derived from an adequate transformation of the pixels of the lip region of interest. Two most popular approaches includes: Geometric feature-based approach and appearance-based approach. Geometric feature-based

approach obtains geometric information of the lip such as the height or weight as features while appearance-based approach takes all the pixels within the mouth region and applies dimension reduction technique such as PCA.

4. **Obtain Sub-visemic classes**:

Visemic is the basic visual unit for speech. We apply clustering such decision tree to group a sequence of images and find each visemic.

5. **Speech Classifier**:

We build classification models through training samples. We try multiple classifiers such as Artificial Neural Network (ANN), Hidden Markov Model (HMM) and Support Vector Machine (SVM).

III. MILESTONES

10.23	Iteration on proposal, and finalize idea
11.1	Implementation detection for face and ROI
11.15	lip shape estimation and Visual feature extraction
11.25	Progress report
12.1	speech classification for digit
12.2-4	Group presentation
12.16	Final report

IV. References

[1] Hassant A.B.A.. (2011). "Visual Speech Recognition. *In: Ivo Ipsic Speech Technologies / Book 2. Rijeka: InTech - Open Access Publisher*. ISBN: 978-953-307-322-4
link : http://cdn.intechopen.com/pdfs/16013.pdf

[2] Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech.*Proceedings of the IEEE*, *91*(9), 1306-1326.
link: http://www.ifp.illinois.edu/~ashutosh/papers/IEEE%20AVSR.pdf

[3] Movellan, J. R. (1995). Visual speech recognition with stochastic networks.*Advances in neural information processing systems*, 851-858.
link: http://mplab.ucsd.edu/wp-content/uploads/movellan94.pdf