Electronic Electronic Letters on Computer Vision and Image Analysis 15(2):30-32, 2016

Speech Recognition Supported by Lip Analysis

Waqqas ur Rehman Butt*

* *University of Pavia, Department of Electronics, Computer Science and Electrical Engineering, Pavia,Italy*

Advisors: Prof. Virginio Cantoni and Prof. Luca Lombardi
Date and location of PhD thesis defense: December 21, 2015, University of Pavia
Received 30 June 2016; accepted 30 July 2016

Abstract

Computers have become more pervasive than ever with a wide range of devices and multiple ways of interaction. Traditional ways of human computer interaction using keyboards, mice and display monitors are being replaced by more natural modes such as speech, touch, and gesture. The continuous progress of technology brings to an irreversible change of paradigms of interaction between human and machine. They are now used in daily life in many devices that have revolutionized the way users interact with machines. In fact new PCs, tablets and smartphones are moving increasingly toward a direction that will bring in a short time to have interaction paradigms so advanced that will be completely transparent to users. The various modes of human-machine interaction, through voice recognition are without doubt one of the most considered.

A number of researchers have revealed that a speech reading system is beneficial complement to an audio speech recognition system by using of visual cues of the speakers, such as face in noisy environment. However, robust and precise extraction of visual features is a challenging problem in object recognition, due to high variation in pose, lighting and facial makeup. Most of the existing approaches use constraints such as the use of reflective marker on subjects lips, lip movements recorded with a fixed camera position (head mounted camera) and lip segmentation in organized illumination conditions. Furthermore, there is no common consensus about the visual features selection and their significance for a particular phoneme.

Speech is the natural procedure of communication. Therefore speech would be an apparently preferred option for human computer interaction. In the past years, development in technology, combined with a significant reduction in cost, has led to the pervasive use of automated speech recognition in variety of systems such as telephony, human-computer interaction and robotics.

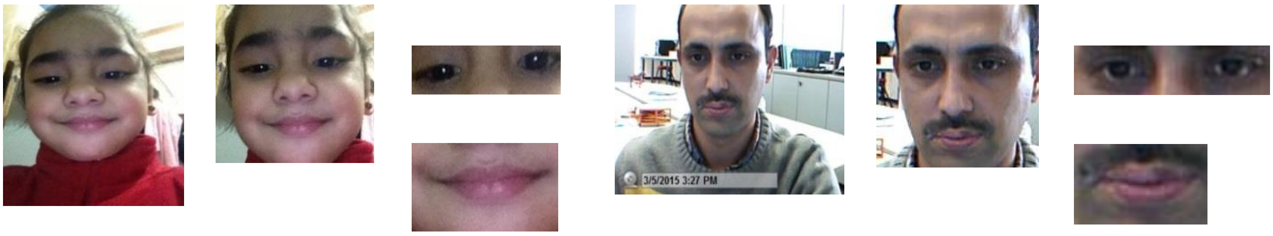
Visual speech cues are prospective source of speech information and they are apparently not affected in noisy acoustic environmental condition and cross talking between speakers. Visual information of a speaker is the key component of Speech Recognition system such as outside area of mouth, mouth gestures and facial expressions.

The major problem to develop robust speech recognition system is to find the precise visual feature extraction method. Sometime hearer observes improper from speaker because of the incompatible effect of visual features. These visual features have great role in the lip reading process. These interpretations gave a motivation for developing a computer speech recognition system.

Correspondence to: wkbutt@hotmail.com

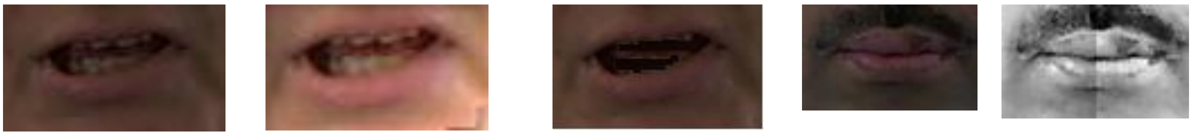
Recommended for acceptance by David Vázquez Bermúdez
DOI: http://dx.doi.org/10.5565/rev/elcvia.953
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

I propose a speech recognition system using face detection, lip extraction and tracking with some pre-processing techniques to overwhelmed the pose/lighting variation problems. The proposed approach is useful for face/lip detection and tracking in sequence of images and to augment global facial features to improve the recognition performance.



a)Original Image a1)face detection a2) eyes and Face b)Original Image b1)face detection b2) Eyes and Face

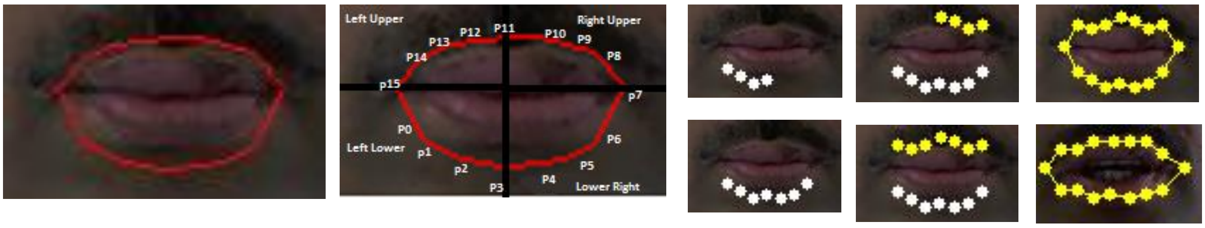
Figure 1. Face, eyes and mouth detection from videos



a) Original b)After Illumination c) Teeth filtering d)Original e)Shadow Filtering

Figure 2. Some results of Illumination equalization, Teeth and Shadow FIltering

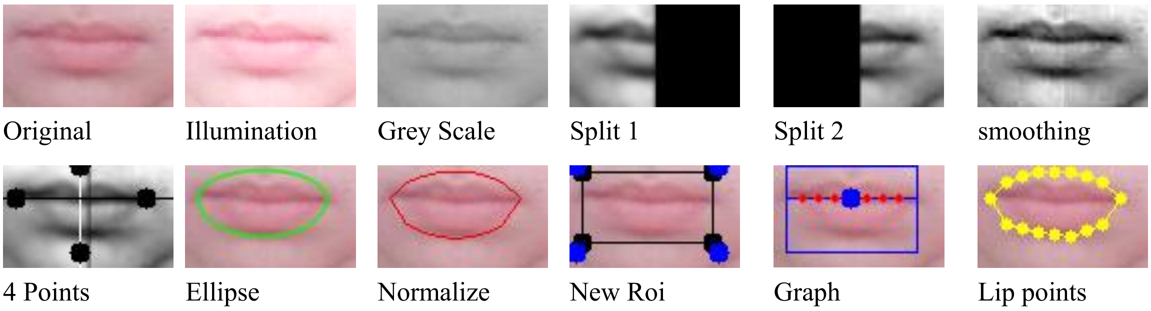
The Proposed approach consists of four major parts, firstly detecting/localizing human faces, lips and define the lip region of interest in the first frame as shown in Figure 1, secondly three pre-processing steps, namely illumination equalization, teeth detection and shadow removal developed, aiming at investigating edge information and global statistical characteristics which are sensitive to the uneven illuminations and susceptible to the complex appearance in presence of teeth and shadow. In contrast, the proposed method, which is aimed at local region analysis, can successfully avoid the complex appearance (e.g. low contrast, shadow, moustaches and teeth). The high average extraction performance is reached as shown in Figure 2, thirdly create contour line (3a), draw the 16 points by splitting image into four parts as shown in the figure 3 (b), and stored the coordinates of these constraints. The new approach is implemented in the lip tracking module. Using this lip tracking module from the lip boundary lines a feature vector of 16 points lip model of the speaker's lips, stores the coordinates of these points and tracks these coordinates during the utterance by the speaker and tracked in every image of the image sequence.



Contour line 16 Points Draw Points d)Open/Close

Figure 3. (a) Contourline (b) Drawn points of the four parts of the image

Finally track the lip contour with their coordinates in the following frames. Extensive experiments show the encouraging results and the effectiveness of the proposed method in comparison with the existing methods. The proposed approach has also been evaluated by testing the system in noisy real world facial image sequences. Experiments have shown that outliers detecting and better predicting ROIs can further reduce the number of frames with locating or tracking failures. Figure 4 shows the complete proceedure for speech recognition system on first frame of the video.



Original Illumination Grey Scale Split 1 Split 2 smoothing

4 Points Ellipse Normalize New Roi Graph Lip points

Figure 4. Complete results for single image by proposed SRS

References

[1] Waqqas ur Rehman Butt, Luca Lombardi "*An Improved Local Region Based Approach for Lip Detection and Tracking towards Speech Recognition*" submitted to Journal of Visual Communication and Image Representation, Under review (June 2016)

[2] Waqqas ur Rehman Butt, Luca Lombardi, Dr. Marion Pause "*Automatic Object detection in digital Images under non standardized Conditions*" submitted to Signal, Image and Video Processing Journal, Under Review (2016)

[3] Waqqas ur Rehman Butt, Luca Lombardi "*A Methodological Comparison of Moving Object Detection and Tracking in Videos with Background Subtraction and Mixture of Gaussian* ". accepted a regular research paper (RRP) in (IPCV'15)The 2015 International Conference on Image Processing, Computer Vision & Pattern Recognition (July 27-30, 2015, Las Vegas, USA)

[4] Luca Lombardi, Waqqas ur Rehman Butt, Marco Grecuccio "*Lip Tracking Towards An Automatic Lip Reading Approach*". Journal of Multimedia Processing and Technologies Volume 5 Number 1 March 2014, Pages 1- 11 , Print ISSN: 0976-4127, Online ISSN: 0976-4135

[5] W.U.R. Butt, L. Lombardi "*Comparisons of Visual Features Extraction Towards Automatic Lip Reading*" 5th International Conference on Education and New Learning Technologies, Barcelona, Spain. (1-3 July, 2013) EDULEARN13 Proceedings, Pages: 2188-2196, ISBN: 978-84-616-3822-2, ISSN: 2340-1117. http://library.iated.org/view/BUTT2013COM