Report: Hons. Project

Shubham Rathore

201430101

Abstract

An efficient approach for speaker detection has been introduced in this project. The project aims to recognize the persons present in the scene and determine the speaker in the conversation with the help of visual information that is supported by audio information.

1 Introduction

The first part of this project aims to detect the speakers present in the scene. The speaker detection is based on lip movements which is supported by Voice Activity Detector(VAD). The receiver detection is based on assumptions that are highlighted below in the report. Histogram processing and Haar cascades are the primary components used.

2 Details

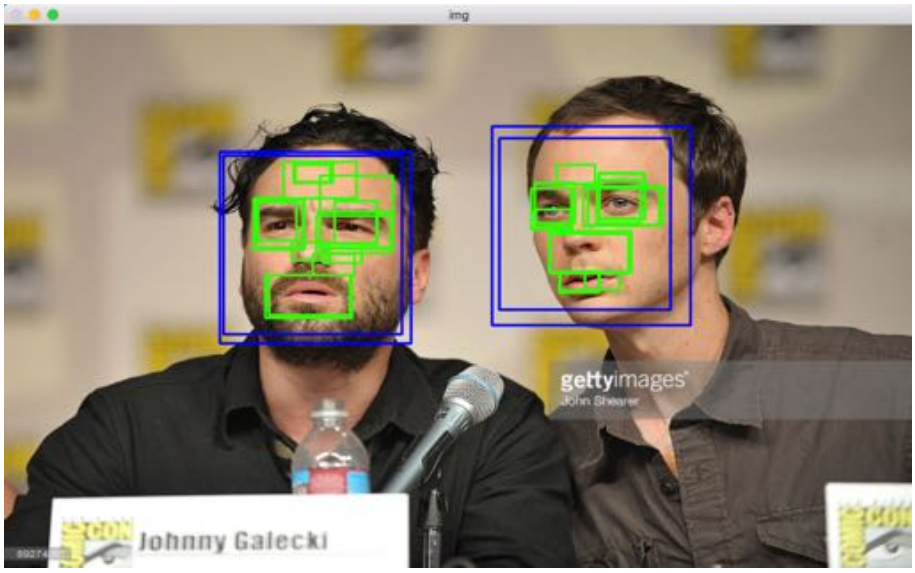
Input: Target video from which the frames are processed
Output: Labels corresponding to the speaker and the receiver
Implementation: OpenCv, Python

3 Process flow

The project comprises of face detection and face recognition thereafter. Then comes the lip movement detector part, that helps in determining the speaker at that moment. Voice Activity Detection(VAD) is used to support the speaker detection part. Finally, the receiver is detected based on assumptions, as mentioned below.

3.1 Face Detection

Face Detection is the primary part that lays the foundation for the entire project. Haar features [5] are used to accomplish this task.It is a machine learning based approach where a cascade function is trained from a lot of positive and negative images. These features work only on the image intensities. A cascade of classifiers is used in which we discard a feature if it fails at the first stage. The cascade used comprises of front and side face cascades as well as a cascade for frontal face with glasses. This way all the faces present in the frames are detected.



3.2 Classification

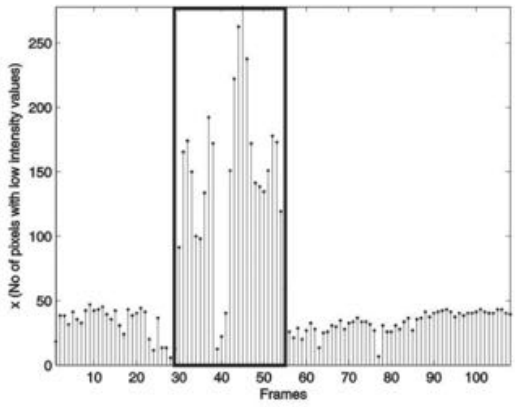
The faces found using the haar features are extracted from the frames and thus a dataset comprising of the faces and their corresponding labels is created. Local Binary Patterns Histograms algorithm is used for face recognition. The basic idea of Local Binary Patterns is to summarize the local structure in an image by comparing each pixel with its neighborhood. Take a pixel as center and threshold it against its neighbors. If the intensity of the center pixel is greater-equal its neighbor, then denote it with 1 and 0 if not. You'll end up with a binary number for each pixel, just like 11001111. So with 8 surrounding pixels you'll end up with 256 possible combinations, called Local Binary Patterns. This forms a basis for classification and hence image recognition is performed.

3.3 Speaker detection

Speaker can only be detected using lip movements when it comes to visual features. The lip detection is localized to the faces detected rather than the entire frame. Again Haar Cascade corresponding to the lip detection is used. Signal detection algorithms are applied to the features of the mouth region. It is observed that the standard deviation of the number of pixels with low intensities in the mouth region can be exploited.[4] The number of pixels with low intensities in the mouth region of a speaking person can be used as visual cues for detecting visual speech. We employ the lip activity detection method in order to determine the active speakers in an environment with multiple persons.

3.4 Histogram Processing

The number of low intensity pixels increase in the mouth region when the mouth is opened. This is due to the exposure of a part of the interior of the mouth, which is usually in shade. Majority of the words spoken involve an open mouth, thus this algorithm works in all cases with a good efficiency. The percentage of low intensity pixels keep on changing drastically when a person is speaking. Pronunciation of phonemes decides the percentage of the mouth open, and thus there occurs a large deviation in the corresponding region of the histogram. On the other hand, when there is no lip activity i.e. no speech the lips are most probably closed and, thus there happens to be no increase in the low intensities of the mouth region and no fluctuation of these intensities. We, thus, argue that the increase and the fluctuation of the number of mouth region pixels exhibiting low intensity values can indicate lip activity. This principle is used in this project for the visual detection of speech.

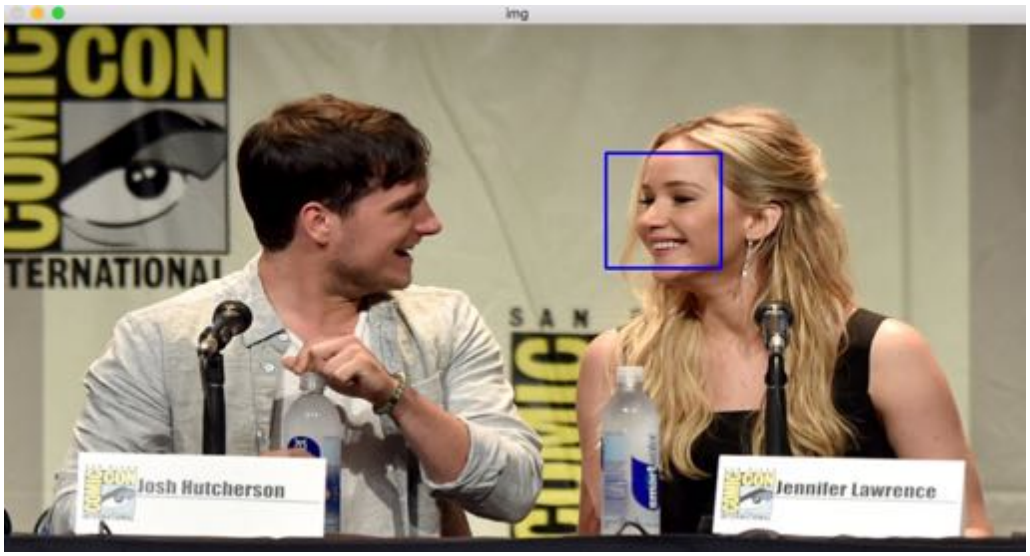


3.5 Voice Activity Detector

Voice activity detection (VAD) is basically the detection of the presence of human speech in an audio clip.[1] The purpose of VAD in this project is to support the speaker detection algorithm that only uses visual information. The inclusion of audio information helps in improving efficiency, as the code for histogram processing is only executed when a boolean true is returned by the VAD system. A boolean true indicates the presence of human speech. The implementation includes the use of numpy and scipy python libraries.

3.6 Receiver Detection

This part of the project is based on assumptions. In simple words, efficiency obtained is quite less. This is simply based on the fact that the direction of the face of the speaker determines the receiver on the other end. The sideface haar cascade is used to determine the direction of the face of the speaker. The video on which the algorithms are tested is a video in which speakers sit next to each other in a parallel manner, thus prediction of the receiver becomes a hectic task, with less accuracy.



4 Further Work

Accuracy of speaker detection can be improved by performing some tweaks in the algorithm proposed. Gender classification can be done[3]. The Voice Activity Detector(VAD) is capable of of distinguishing male versus female speech[2]. This way we can narrow down our search for the speaker which will in turn help in attaining a greater efficiency. The mouth detection and the histogram processing algorithm will only be applied to the faces according to the gender detected in the face detection and the VAD part of the process.

References

[1] Fasih Haider and Samer Al Moubayed. Towards speaker detection using lips movements for humanmachine multiparty dialogue. *FONETIK 2012*, page 117, 2012.

[2] Hadi Harb and Liming Chen. Gender identification using a general audio classifier. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 2, pages II–733. IEEE, 2003.

[3] Vladimir Khryashchev, A Priorov, L Shmaglit, and M Golubev. Gender recognition via face area analysis. In *World congress on engineering and computer science*, 2012.

[4] Spyridon Siatras, Nikos Nikolaidis, Michail Krinidis, and Ioannis Pitas. Visual lip activity detection and speaker detection using mouth region intensities. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(1):133–137, 2009.

[5] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.