Lip-Contour based Speaker Activity Detection in smart environment

Jagdish Lal Raheja[1], Surabhi Sharma[2,]
CSIR-Central Electronics Engineering Research Institute
(CSIR-CEERI), Pilani 333031, Rajasthan
[2]PEC University of Technology, Chandigarh 160012

Axel Plinge[3], Gernot A Fink[3]
[3]Intelligent Systems Group,
Robotics Research Institute, Technical University
Dortmund, Germany

Abstract— **In this paper we describe a lucid model for efficient localization of an active speaker in a real-time video for a video conference. The system is built on a simple and robust method of contour detection of grey images and frame differencing of binary images for motion detection. This proposed method is independent of luminosity and removes unnecessary parameters from the frame to be worked upon so that the processing time is reduced and a faster algorithm is obtained to detect the speaking person. We aim at developing a robust technique for talker identification by processing a continuous video input from a live web camera and removing the necessity of using the cumbersome process of identifying a speaker in a discussion or a video-conference and then focusing on the person manually. Results of testing the suggested model are presented with conclusions.**

Keywords- **Speaker, Lips**, ***Contour Detection, Frame Differencing***

1. INTRODUCTION

Detecting lip movement and identification of the speaker [1] are areas of active research in video surveillance and human-computer interaction. [2]. These techniques are a motivation of the rapidly growing audio-visual speech recognition inspired by "Hearing Lips and Seeing Voices" [3]. Basically, in a video conference, the speaking person can be identified and the camera can be focused on the respective person automatically rather than doing it manually. For accomplishment of this task we must be able to extract the human faces from the video stream with accuracy, detect the lips in the faces and then process each in order to localize the speaker. There had been different approaches to face detection, lip detection and lip movement detection in recent past [4] but rely on computational expensive method called Coherent Point Drift (CPD) to detect the lips movement. Face detection based on skin filters[4],[5],[6],[7], neural networks [7],[8],[9], genetic algorithms[10], etc. have been done but that requires proper background subtraction and poses problems in case "certain background creeps in" while the color thresholding is being done.

Various methods for lip contour detection have been implemented earlier which include use of sobel filter[11],[12], active contours[13], cubic spline interpolation[18], color thresholding[4],[5],[6],[7],[14], having a database of lip images[15],[16],[17], etc. The problem with these methods is that they vary with the luminosity variations, make-up conditions, blemishes on the skin, etc. Movement of lips has been detected through Kalman filter and optical flow [15],[18],[19]. Optical flow is prone to rotation of the object and illumination. Kalman filter needs a lot of processing time and memory space.

The probability based method for speaker identification in [20] is cumbersome. It associates two probabilities with each face, one with identification in the video frame and another with speaking as detected through the amalgamation of audio and video input. It's gives undesired results in case of short utterance detection, departure or arrival of a speaker randomly and does not have high speaker change detection.

Some of the earlier developments on the same lines do not take into account more than one face in a frame and thus are not suited for real time systems.

In this paper, we aim at ameliorating a model which will cater to the shortcomings in previously employed methods and develop an integrated system for speaker localization in video conferencing. It is based on a simple logical analysis of the frames from a video stream to continuously monitor the participants and notify whenever a person is witnessed as a speaker and the speaker in shown separately in the broadcasting window.

2. SYSTEM OVERVIEW

The flowchart for the proposed model is shown in Fig.1. In this model, we simply define a model where a human face is identified using change in contrast values between adjacent rectangular groups of pixels via Voila-Jones algorithm. Lip contour is detected with edge detector and their movement is distinguished through frame differencing. Thus, this is done for all the faces in the frame and the active speaker is localized.

The following sections will expatiate on the proposed idea in further details.

3. PROPOSED METHOD

The proposed method implies to the requisite of luminosity independent method and a completely automated active speaker localization system. The gradual development of the method is being discussed piecemeal.

3.1. FACE DETECTION

The first and foremost step is to detect faces from a continuous video-stream with a complex background.
The code must be robust enough to detect face from the video with flimsiest error in detection because it will foster the process of probable lips location and sleuthing of the movement in order to detect the one who is speaking.

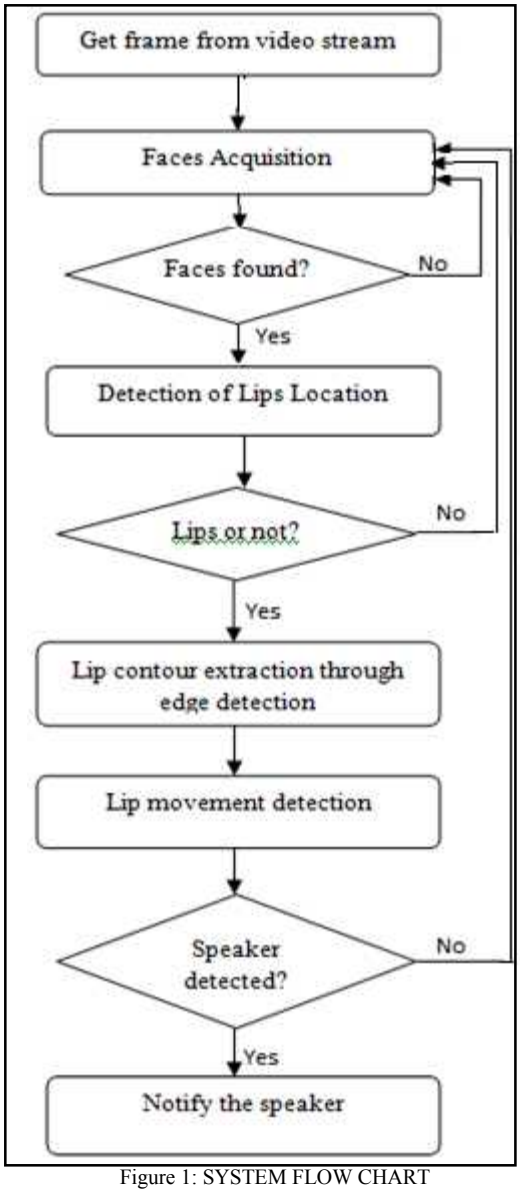


Figure 1: SYSTEM FLOW CHART

Irrespective of the method which includes the usage of skin-filters[4],[6],[8], [21] we use the Haar-like classifier [5] ,[20] i.e. the Viola-Jones Algorithm[22] to detect human faces in the input video stream.

3.1.1. BASIC LOGIC

A recognition process can be much more efficient if it is based on the detection of features that encode some information about the class to be detected. Well know Viola-Jones algorithm [22] is used to detect face as shown in fig. 2. Detected face is shown by a red color rectangular box in the frame. It overcomes the cumbrous chore of detecting faces through HMM method, deformable templates [17], skin filter[4] [5] , etc.

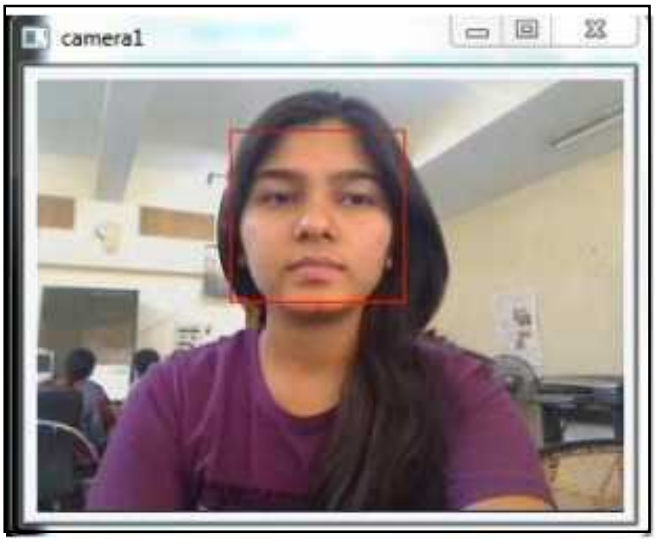


Figure 2: Face detected through Viola Jones algorithm

The method is very robust against varying lightning conditions. Occasionally, face-like structures in complex backgrounds lead to false positives. To exclude these, the additional constraint that the region contains mainly skin color is introduced. This is done by transforming the face rectangle into Y, Cr, Cb-space [2] and counting the pixels within the range $(50 \ldots 170, 70 \ldots 170, 140 \ldots 170)^T$. Detections where the pixel count is below 40% of the face rectangles' area are discarded.

$$n_{skin} \geq 0.4 \cdot h_f \cdot w_f \qquad (1)$$

3.1 Lip Contours

It was found that speech activity is reflected by a significant change in the lip contour of the speaker. This can be detected by a fast online algorithm counting the number of pixels in the normalized lip countour in consecutive face detections. First, the mouth region is computed from the face region as

$$x_m = 0.90(x_f + 0.35w_f) \quad (2)$$
$$y_m = 3.00(y_f + 0.22h_f)$$
$$h_m = 0.38h_f$$
$$w_m = 0.43h_f \ .$$

By edge detection, the lip contour $l_{i,k}$ is derived and the number of contour pixles l is counted for each face i and time frame k. Speaker activity is determined by comparing the lip contour size in consecutive frames. For each face detection, the relative number l of lip contour pixels for face i in time frame k is determined

$$l_{i,k} = \frac{100 \cdot c_{i,k}}{h_{mouth} \cdot w_{mouth}} \qquad (3)$$

A speaker is considered actively talking if the lip contour sizes in frames within a time window of th_d differ more than a threshold th_l.

$$dl_{i,k} = \max \ |l_{i,k} - l_{i,k-d}| \ \ 0 < d/fps \leq th_d \quad (4)$$
$$vad_{i,k} = dl_{i,k} > th_l \quad (5)$$

4. EXPERIENTAL RESULTS

In section 3, the whole algorithm for lips motion detection has been discussed. The code was enhanced so as to make it more user-friendly and the steps for the same have been described here under.

Firstly, whenever a single person spoke in a camera the Command Prompt depicts the motion by showing the text "Lips Moving". In case the person doesn't speak, no notification is given. A snapshot of the same can be seen in Fig.3.

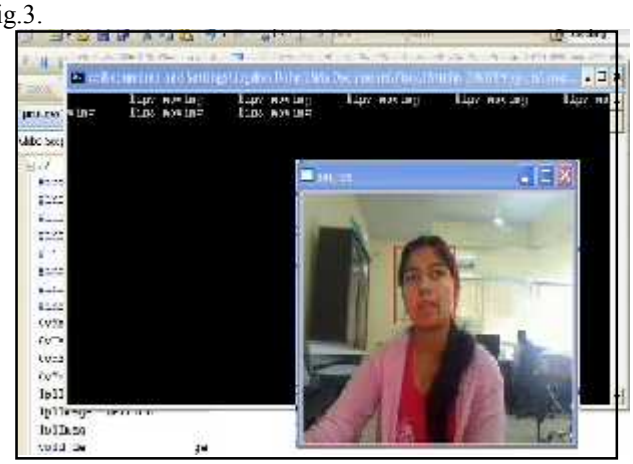


Figure 3: Speaker Localization through 1 camera.

After completing lip movement detection through one camera for just one person in a camera, the code was modified to detect speaking persons simultaneously in two cameras. Two cameras were attached with one person sitting in front of each camera. If the person in front of camera 1 speaks, the code shows- "Lips moving in camera 1". If the person in front of second camera speaks, the code shows- "Lips moving in camera 2". In case, none of the people speak. No notification is seen. A snapshot of the same can be seen in Fig.4.

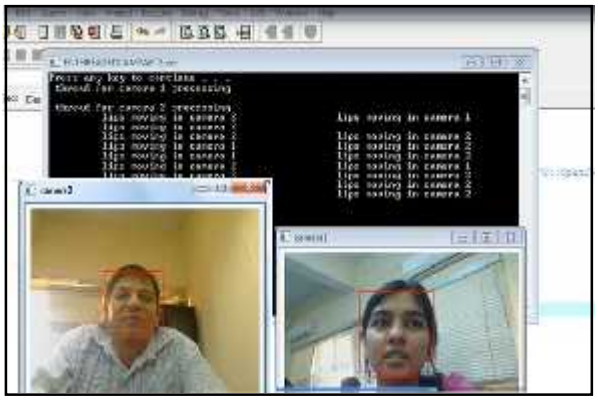
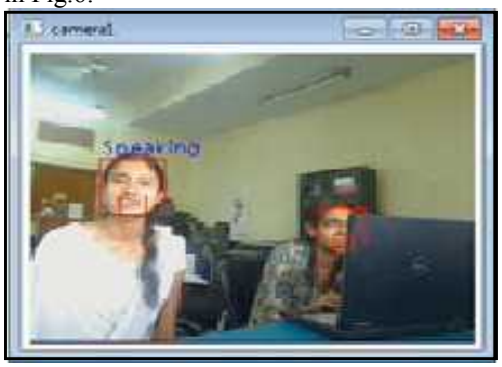


Figure 4: Speaker Localization through 2 cameras.

Successful implementations of simultaneously working two cameras for speaker detection lead to enhancement of code for making it work for four cameras. Whenever a speaking person was identified, it was reported by the code with the camera number thus helping to detect in which camera a person is speaking. The snapshot its implementation is shown in Fig.5.

Figure 5: Speaker localization through 4 cameras.

The code was enhanced further to detect any number of speaking people in any number of camera and the speaking person is identified with the text-"Speaking" being written above the speaking person. This was tested for three cameras with two people sitting in front of each. The results obtained were satisfactory. A snapshot of the execution is shown in Fig.6.



(a)

(b)

(c)

Figure 6: Speaking people detected in camera 1 and camera 2

4.1. QUANTITATIVE ANALYSIS

In order to derive quantitative results, a sequence was analyzed in detail. In one session, a single speaker was facing the camera, alternating speaking and silent while looking into the camera as shown in fig. 7. It was recorded at 30 fps. Figure 8 illustrates the detection (a) speech activity as function of the parameters th_l and th_d. Window sizes above 0.5s yield a good recall. Thus the window size was fixed here to yield good results with a small delay. Figure 9shows the influence of th_l. It was set to 0.8 to give a good precision/recall tradeoff.

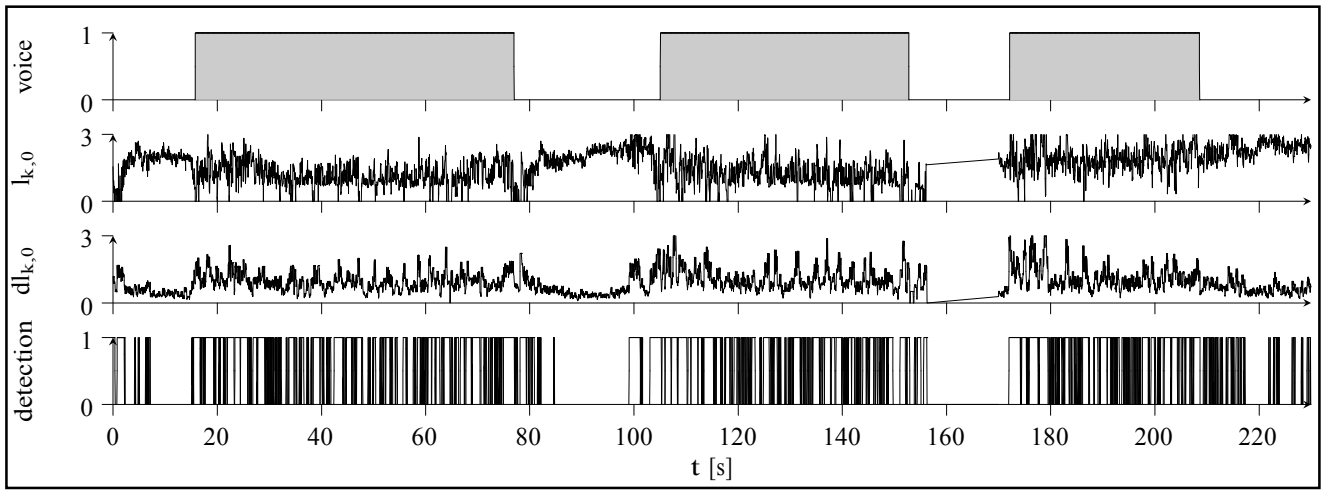


Figure. 7. Ground truth (top), lip contour sizes, contour delta and lip-based speech detection

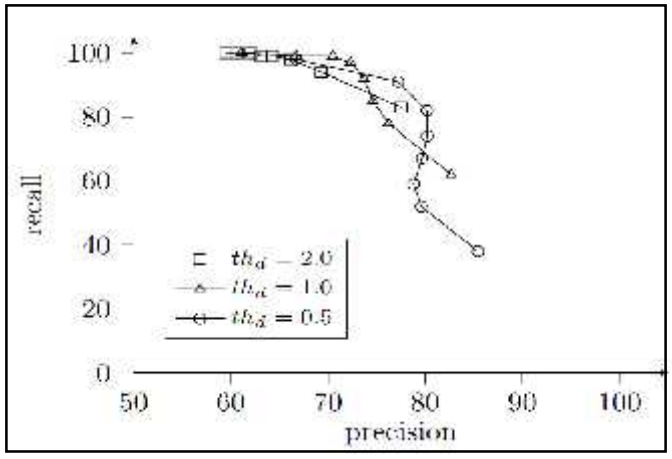


Figure 8: The ROC diagram shows curves for different th_d. The curves themselves are parameterized by th_l

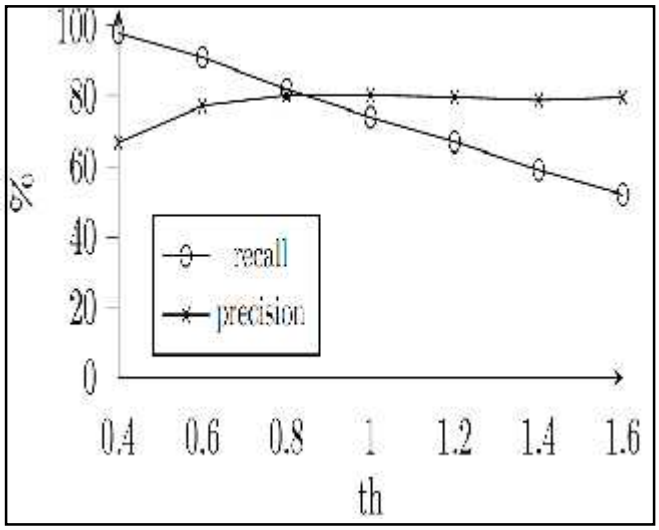


Figure 9: Precision and recall for $th_d = 0.5s$

seqence	Length	fps	F1	Precision	Recall
Seq02	248s	20	75.57	83.52	69.0
Seq06	62s	20	89.61	94.75	85.0
Seq07	101s	20	81.09	80.21	82.0

Table1: Results for threee recordings of speakers facing the camera

Several recordings of speakers facing various cameras were made to verify the method. Three recordings of single speakers facing the camera were made. They did altenate talking and not talking over the course of the recording. In table 1, the results are given. Figure 7 illustrates the speaker activity results for the first sequence (sq02).

5. CONCLUSIONS

A unique and robust real time algorithm, which takes up the task of detecting the movement of the lips, has been proposed. The proposed technique is independent of the scale of the face image and performs well irrespective of the luminosity variations. Some of the erroneous results were given due to reasons like- input stream of images from live web camera being too blurred, extreme complexity of background, movement of face, moustaches and beard of the person. To handle these problems, a tracking method can be integrated and audio based voice activity detection (vad) could be integrated.

ACKNOWLEDGMENT

The work has been jointly supported by the CSIR, India and DAAD, Germany and is highly appreciable. Further authors would like to thank Director, CEERI, India for his valuable advice and Director, TU Dortmund, Germany, and his department for providing the infrastructure and technical support.

REFERENCES

[1]. Cha Zhang, Yong Rui, Li-wei He, "Hybrid Speaker Tracking in An Automated Lecture Room". Communication and Collatoration Systems Group Microsoft Research One Microsoft Way, Redmond, WA 98052.

[2]. Pantic M., Rothkrantz L.J.M., "Toward an affect-sensitivemultimodal Human-Computer Interaction". Proceedings of the IEEE, Volume.91, Issue.9, pp. 1370–1390, 2003.

[3]. Harry McGurk, John MacDonald, "Hearing lips and seeing voices". Nature 264, 746-748, 23 Dec 1976.

[4]. Saravi, S. Kalawsky, Roy S. Zafar, Iffat Edirisinghe, Eran A., "Speaker identification Coherent point drift Lip movement detection", Proc. SPIE 7724, Real-Time Image and Video Processing 2010, 77240D (May 04, 2010)

[5]. Usman Saeed, Jean-Luc Dugelay, "Facial Video Based Response Registration System". 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, Copyright by EURASIP.

[6]. Xiaozheng Zhang and Charles C. Broun, "Using Lip Features For Multimodal Speaker Verification". 2001: A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001.

[7]. Y. Mitsukura, M. Fukumi and N. Akamatsu, "A Design of Face Detection System by using Lip Detection Neural Network and Skin Distinction Neural Network". pp-2789-2793, 2000 IEEE.

[8]. H.A. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," IEEE Transactions Pattern Analysis and Machine Intelligence, pp.23-38, January 1998.

[9]. L. Mostafa and S Abdelazeem, "Face Detection Based on Skin Color Using Neural Networks", GVIP 05 Conference, Cairo, Egypt, Dec 2005.

[10]. H.Yokoo, M.Hagiwara, "Human Face Detection Method Using Genetic Algorithm". "The Journal of The Institute of Electrical Engineers of Japan (Electronics, Information and Systems Society), Vo1.117, No.9, pp.1245-1252(1997) in Japanese.

[11]. Siew Wen Chin , Li-Minn Ang, Kah Phooi Seng, "Lips Detection for Audio-Visual Speech Recognition System". 2008, IEEE.

[12]. Koji Iwano, Tomoaki Yoshinaga, Satoshi Tamura, and Sadaoki Furui, "Audio-Visual Speech Recognition Using Lip Information Extracted from Side-Face Images". EURASIP Journal on Audio, Speech, and Music Processing Volume 2007, Article ID 64506, 9 pages doi:10.1155/2007/64506, 2007.

[13]. Chan, T.F., Vese, and L.A, "Active contours without edges". IEEE Transactions on Image Processing, vol.10, no.2, pp.266-277 (2001).

[14]. Muraselet al, "Fast Visual Using Focused Color Matching-Active Search, IEICE Tech. Report (D-11), Vol. J81-D-11, No.9, pp.2035-2042(1998) in Japanese.

[15]. Yaser Yacoob and Larry S. Davis, "Recognizing Human Facial Expressions From Long Image Sequences Using Optical Flow". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.18, pp.636-642, No.6, June 1996.

[16]. Robert W. Frischholz, Ulrich Dieckmann, "BioID: A Multimodal Biometric Identification System". IEEE, February 2000.

[17]. A.L. Yuille, D.S. Cohen and P.W. Hallinan "Feature Extraction from Faces Using Deformable Templates". IEEE Conference Computer Vision and Pattern Recognition, pp.104-109, 1989.

[18]. K. Iwano S. Furui T. Yoshinaga, S. Tamura, "Audio-visual speech recognition using lip movement extracted from side-face images," Proc. Auditory Visual Speech Processing (AVSP), pp. 117–120, 2003

[19]. S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical flow analysis for lip images". Journal of VLSI Signal Processing, Volume.36, Issue.2, pp.117–124, 2004.

[20]. Athanasios Noulas, Gwenn Englebienne, and Ben J.A. Kröse, "Multimodal Speaker Diarization". IEEE Transactions on Pattern Analysis and Machine Intelligence, 10.1109/TPAMI.2011.47, 2011
.

[21]. Kshitiz Kumar, Tsuhan Chen and Richard M. Stern, "Profile View Lip Reading". *IEEE* Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, 2007.

[22]. Gary Bradski & Adrian Kaehler, "Learning OpenCV-Computer Vision with the OpenCV Library". Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.