Novel Pixel-based Approach for Mouth Localization

P.Sujatha ,
Department of CSE, Sudharsan Engineering College, Pudukkottai, Tamilnadu.

M.Radhakrishnan, Ph.D
Director / IT, Sudharsan Engineering College, Pudukkottai, Tamilnadu.

ABSTRACT

Mouth localization is used in many applications such as face detection and lips reading. Visual information from lip movements helps to improve the accuracy and robustness of a speech recognition system. This paper presents a new method for automatic lip detection using geometric projection method and adaptive thresholding. From the real time video, the face images are grabbed and a modified geometric projection method is proposed to extract the mouth region based on the distribution relationship with the face Region Of Interest (ROI). After mouth localization, a new pixel-based approach is proposed to extract the outer lip contours. The performance of the lip tracking method using adaptive thresholding is evaluated in real time in the normal room environment, and this method achieves 98% recognition rate.

Keywords

Mouth localization, Geometric projection method, Lip tracking, Adaptive thresholding.

1. INTRODUCTION

Several researches have demonstrated that useful information about the speech content can be obtained through lip reading of speakers [1-5]. Lip reading is a technique of understanding speech by visually interpreting the movements of speaker's lips. Lip localization is the first step in lip reading system and if it is not accurate it directly affects the lip tracking and feature extraction of lip movement and ultimately will have an impact on the recognition rate [1]. The main goal of lip reading research is to make the human-computer interaction more natural and to adapt with different lighting conditions, different speakers and also with various skin colors. There are wide ranges of application in which lip reading is an integral part that can improve the performance of overall system. These applications include audio-visual speech recognition (AVSR), visual speech recognition (VSR), synthetic talking faces and facial expression analysis. Currently, significant research efforts are being made on AVSR and VSR. AVSR is the extension of acoustic speech recognition, which employs both acoustic and visual information. It significantly improves the recognition accuracy in noisy environments [2]. Visual speech recognition is a vision based approach to recognize speech without evaluating the acoustic signal. Potential application of such a system includes human computer interface for hearing impaired users, lip reading mobile phones and improvement of speech-based computer control in noisy environments [3]. Difficulties in the audio based speech recognition system can be significantly reduced by additional information provided by the extra visual features. It is well known that visual speech information through lip-reading is very useful for human speech perceptions.

The aim of this paper is the extraction of inner and outer lip contour using geometric projection method and adaptive thresholding. This paper is organized as follows. Section 2 gives the technique used in this paper for face localization. Section 3 explains the geometric projection method for lip localization. Section 4 describes the pixel-based technique used in this work for lip tracking. Section 5 presents the experimental results and section 6 presents conclusion.

2. REAL TIME FACE LOCALIZATION

Most of the researchers have done their work starting with face detection and then later on lip localization [1-2]. Face detection is used to find the faces in the given arbitrary image. Different techniques have been proposed for face detection and localization. In this paper, Viola and Jones face detector [6] is used. This method is capable of processing image promptly and achieving high detection rates. The Viola and Jones face detector method has been distinguished by three key contributions such as new image representation called integral image, learning algorithm based on AdaBoost and combining classifiers using a cascade scheme. The first contribution in face detection is the new image representation called integral image representation. It allows the features used by the detector to be computed very quickly. For each pixel in the original image, there is exactly one pixel in the integral image, whose value is the sum of the original image values in the left and above of the original pixel.

$$I\_I(x, y) = \sum_{x' \le x,\, y' \le y} I(x', y') \qquad (1)$$

In (1), $I\_I(x, y)$ is the integral image at location x, y and $I(x, y)$ is the original image.

The second contribution was an AdaBoost learning algorithm which selects a small set of features from a large set and yields extremely efficient classifiers. Given a set of positive and negative training image features, the AdaBoost classifier is employed to boost the performance of a weak classifier.

The third contribution is a method for combining complex classifiers in a cascade to increase detection performance while reducing computation time. It allows background region of the image to be quickly discarded while spending more computation on promising object-like regions. Majority of the sub-windows are rejected before complex classifiers to achieve low false positive rates

In this paper, in-house audio visual dataset is used. The recording details of the database are briefly explained in section 5. The video is captured in Audio Video Interleave (AVI) file format which is taken as input to face detection module. Then the frames are grabbed from the video and it is subjected to the face detection module as a JPEG (Joint Photographic Experts Group) image. From the input images, large number of features were evaluated using two rectangle, three rectangle and four rectangle features. The two-rectangle and three-rectangle features were overlaid on a typical training face in the video.

From the large number of features, a small set of critical features were selected using Adaboost learning algorithm. Then the critical features were classified by combining complex classifiers. Most of the negative feature sub-windows were rejected before the complex classifiers. In the experiments, about 70 – 80% of the candidates were rejected in the first two stages – feature evaluation and feature selection, which made this technique speed up the detection.

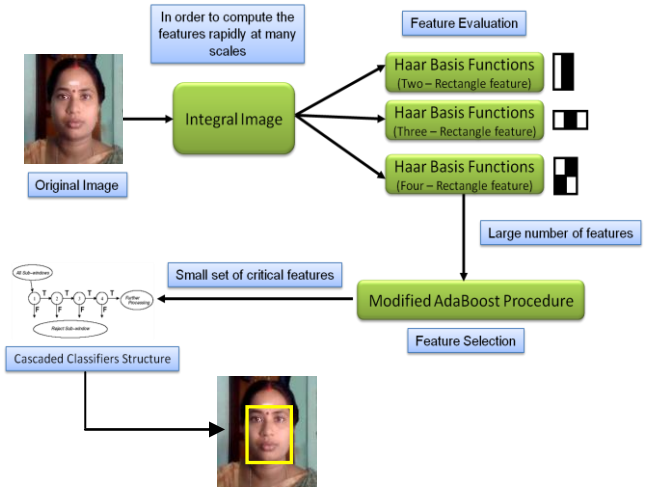

Figure 1: Face Localization process using AdaBoost classifier

Fig. 1 shows the overall face detection module which includes feature evaluation, feature selection and combination of classifiers. In-house database is taken as input to face detection module which detects the face and it is marked by a rectangle ROI. The results after face detection under different lighting conditions, backgrounds, face poses are combined and shown in figure 2.

Figure 2: Output of face detector on a number of test videos.

3. MOUTH LOCALIZATION

For the past few years, many techniques have been proposed in the literature to achieve lip detection. In general, mouth localization is categorized into two methods: *Gray projection method* and *geometric projection method*. In *gray projection method*, an image is projected on horizontal and vertical axes, so the mouth region is defined by valleys of horizontal and vertical curves. In this method, mouth region can be easily defined but with less accuracy, which can be easily affected by bad lighting conditions, low discrimination in lip and skin color. In *geometric projection method*, the mouth Region of Interest (ROI) is roughly located according to distribution features of the mouth in the face region. Advantage of this approach is simple and fast mouth localization. Main drawback is less accuracy for different head poses.

In this paper, a modified geometric projection method is proposed to detect the mouth ROI. After face detection, a mouth region is localized by considering the lower part of the face region defined empirically as follows:

(i) Mouth is always in the lower part of the face.
(ii) Half size of the distance from the face ROI is the width of the mouth.
(iii) 1/3rd of the height of the face ROI is the height of the mouth.

Based on the above analysis, fast mouth detection is proposed using geometric projection method. In the following algorithm, mouth region is located using its distribution relationship with faces.

(1) The result after face localization is taken as input for mouth localization.
(2) Detect the mouth ROI of all the frames of grabbed face ROI.
(3) Find out the values associated with the face region in the x-y coordinates.
face_left: x-coordinate value of the Left border.
face_top: y-coordinate value of the Top border.
face_width: Width value of the face region which is calculated as the difference between x-coordinates of left and right borders of the face rectangle.
face_width = (*face_right_border – face_left*)
face_height: Height value of the face region which is calculated as the difference between y-coordinates of top and bottom borders of the face rectangle.
face_height = (*face_bottom_border – face_top*)

(4) Using the generalized calculations (5-8), the mouth ROI is located in the x-y coordinates of the face ROI.
mouth_width = (face_width+face_left) – (face_width / 4)
mouth_height = (face_height + face_top) + (face_height / 15)
mouth_left = face_left + (face_width / 4)
mouth_top = face_top + (2(face_height /3))*
where,
mouth_width: Width value of the mouth region.
mouth_height: Height value of the mouth region.
mouth_left: x-coordinate value of the left border of mouth region.
mouth_top: y-coordinate value of the top border of mouth region.
(5) *Mouth_width, mouth_height, mouth_left, mouth_top* are the values used to localize and extract the mouth ROI from the grabbed frames of face ROI.
(6) The extracted mouth ROI is copied into new frame for further processing.
(7) Repeat the steps (1) to (6) for all the frames until the video ends.

The diagrammatic representation of mouth ROI extraction using geometric projection algorithm is shown in fig 3. Based upon the rectangle ROI of the face, another ROI is set to extract the mouth in the lower half of the face. Mouth ROI is separated from the frame and it is copied to another frame which has only the mouth region. The proposed method has the advantage of providing a reliable mouth ROI without any mouth model construction and complex procedures such as determining corners and edge detections. This method will be more helpful for those research works which involves the lip reading process.

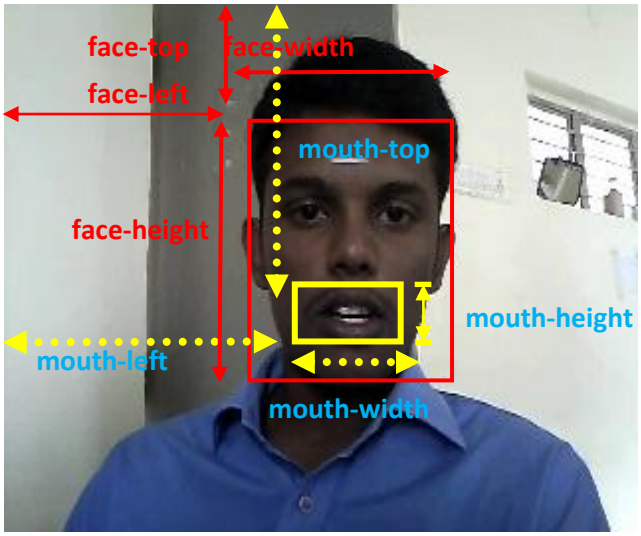

Figure 3: Mouth region localization in real time Video

To extract the lip region, geometric projection based lip detection method is used in this paper. In a standard face the location of the mouth is in the lower half of the face. Based on this concept, a ROI is set by reducing the left, width, top and height values with respect to the face ROI. Then the mouth ROI is localized by the empirical calculations. The extracted Mouth ROI is copied into new frame for further processing. The results after lip localization using geometric projection model are shown in fig. 4.

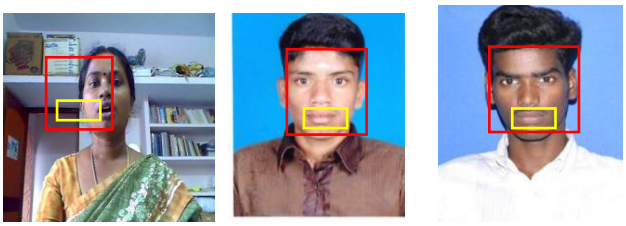

Figure 4: Mouth localization results

4. LIP TRACKIG

After the mouth region localization, a precise lip tracking should be followed for proper lip reading. Lip tracking is challenging because large variations caused by high deformable level of lips, different color tone of lips, illumination conditions, appearance of teeth and tongue, presence of facial hair, beard and so forth. Various methods are available for lip tracking sequences and these approaches can be classified into two major approaches: *the pixel based approach* [7-11] and *model based approach* [12-14]. In *pixel based approaches*, the lip features are directly derived from the given images. The image intensities are pre-processed and then used as a feature vector. Preprocessing normally consists of filtering concepts and reduction in dimension. The advantage of this approach is that there is no data loss and the procedure is easy for deriving the lip features. The disadvantage is it is left to the classifier to learn the nontrivial task of finding the generalization for translation, scaling, rotation, illumination and linguistic variability. Another disadvantage is high dimensionality and high redundancy of feature vector which affects the processing time. Many research works are available based on pixel based approaches.

In *model based approaches*, a model of the visible speech articulators, mainly the lip, is built and its configuration is described by a small set of parameters. The advantage of the model based approach is that the important features are represented in a low dimensional space and are normally invariant to translation, rotation, scaling and illumination. The main difficulty in the model based approach is to build a model which represents the lip shape efficiently and which is able to locate and track the lip contours of different speakers under different illumination conditions. Various research works are there using model based approaches. One of the important issues in model based approach is the formulation of the cost function that drives the lip model to fit the original lip in the image.

In this paper, a new pixel based approach is proposed for real time lip tracking from color images. Accuracy, robustness and processing time are the main concerns of our proposed algorithm. In this paper, lip tracking mainly includes 3 steps: Image enhancement, thresholding and lip contour tracking. After the mouth ROI extraction, the enhancement of the lip region has to be done to yield better result. The enhancement starts from increasing or decreasing the brightness or contrast of the image.

In human perception, brightness is visually judged by the luminance of the object. Brightness is a color coordinate in the HSB or HSV color space (hue, saturation, and brightness or value). By increasing or decreasing brightness we can improve the quality of image. Contrast is another important attribute to improve the quality of the image. Contrast is the difference in visual properties that makes an object distinguishable from other objects and the background. After image enhancement, threshold is used to separate the lip and non-lip region.

Thresholding is used to segment an image by setting all pixels whose intensity values are above a threshold to a foreground value and all the remaining pixels to a background value. It is the simplest way of segmenting an image ROI. Basically there are 3 types of thresholding, which can be used to separate the object from its background. They are global thresholding, local thresholding and adaptive or dynamic thresholding. In global thresholding, the threshold value depends only on f(x, y), where f(x, y) is gray level at pixel (x, y) i.e., intensity value of the x, y coordinates. In local thresholding, threshold value depends on f(x, y) and p(x, y), where p(x, y) is the local property of pixel (x, y) like average gray level of a neighborhood centered on x, y. If the value depends upon f(x, y), p(x, y) and (x, y), where (x, y) is the spatial coordinates of the pixel, it is referred as adaptive or dynamic thresholding. In this paper, adaptive thresholding method is used to track the inner and outer lip contours

The lip and non-lip region can be discriminated using the following algorithm.

1. The frame which has only mouth ROI is subjected to image enhancement

2. The enhanced image serves as the input for thresholding.

3. Adaptive thresholding is applied to the input image.

4. For each pixel in the input image, threshold T is calculated.

5. For all pixels of the lip region, judge their pixel value using the formula:

$$f(x, y) > T \quad (2)$$

where f (x, y) represents the pixel value of (x, y). If it satisfies the equation (11), then it can be considered as lip pixel and set it to black, otherwise non-lip pixel and set it to white.

6. The threshold image is enlarged to the size of 200 * 200 pixels for better processing.

7. The resulting frame after thresholding is a mass of lip contour points where the feature points of inner contour points were extracted for both upper and lower lips.

8. The point of interest (POI) is detected by the projection of final contour on horizontal and vertical axes.

9. The outer contour points are extracted from the first and last gray level changes on horizontal and vertical axes.

10. Repeat the steps from (1) to (9) for all the frames.

Overall, 98% of the lips can be accurately tracked for in-house database. New Pixel based approach using adaptive thresholding have better separation ability comparing to other color space components. Our proposed lip tracking method has successfully improved the lip tracking performance under different lighting conditions, different lip shapes and different lip colors. The lip tracking results are shown in table 1.

Table 1Lip tracking results

	Real time In-house Database
Total No. of mouth ROI frames (25 * 600)	15,000
Detected no. of frames	14,700
Recognition Rate	98%

5. EXPERIMENTAL RESULTS

In order to evaluate the performance of our proposed procedure, in-house dataset is used. The recording details of the dataset are shown in table 3. The in-house videos were recorded inside a normal room using web camera. The participants were 14 females and 16 males, distributed over different age groups, starting from 15 years to 50 years. The videos were recorded at 25 frames per second. It is stored in AVI file format and resized to 320*240 pixels, because it is easier to deal with AVI format and it is faster for training and analysing the videos with smaller frame sizes. Each person in each recorded video utters different phonetically balanced English sentences, which are differing in background, illumination, poses and also in talking style. Ten different subjects were used for 30 persons for which each person utters 10 subjects with 2 different slangs. Thus, this database consists of 30 * 10 * 2 = 600 AVI files. The only restriction on these videos is that they must show the frontal face of the person. Fig. 5 shows the percentage of recognised frames for male and female speakers. Both the speakers pronounced 10 different subjects in two different slangs.

Table 2 In-house database details

	Male	Female
No. of persons	16	14
No. of subjects	10*16 =160	10*14 = 140
No. of times subject is uttered	160*2 = 320	140*2 = 280
No. of frames	320*25 = 8000	280*25 = 7000
Total No. of frames		**15,000 frames**

The frames from the video are subjected to the face detection module which detects the face in the video and marked by a rectangle ROI using AdaBoost cascaded classifier. This real time face recognition was described in section 2. The real time face tracking method is computationally efficient and it is not sensitive to the size of the face, facial expression and lighting condition. Based upon the rectangle ROI of the face, another ROI is set to locate the lip in the lower half of the face as described in section 3. The lip ROI is separated from the frame and is copied to another frame where the current frame has only the lip region.

The frame which has only lip is subjected to image enhancement to improve the quality of image for further processing. The enhanced image serves as the input for adaptive thresholding where lip region is separated from the background. The threshold image is enlarged to the size of 200 * 200 for better results. The resulting frame after thresholding is a mass of lip contour points where the inner and outer contour points were extracted for upper and lower lips. It follows the method described in the section 4. The lip tracking performance of the proposed method with respect to in-house database is given in Table 1.

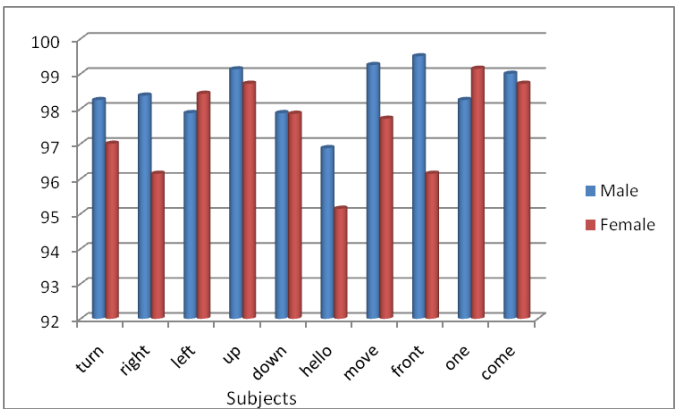

Figure 5: Comparison result of recognized frames percentage for male and female speakers with 10 different subjects.

6. CONCLUSION

In this paper, a new method for lip contour extraction from the face is presented. The recorded visual speech video is given as input to the face localization module for detecting the face ROI. Based upon the AdaBoost cascaded classifier, the rectangle ROI of the face is drawn and another ROI is set to locate the mouth region. The mouth ROI is separated from the frame and is copied to another frame which has only the mouth region. The frame which has only the mouth region is subjected to image enhancement to improve the quality of image for further processing. The enhanced image serves as the input for thresholding, where lip region is separated from the background. The resulting frame after thresholding is a mass of lip contour points where the outer contour points are extracted. The performance of the lip tracking method using adaptive thresholding is evaluated in real time in the normal room environment, and the method achieves 98% recognition rate. This method is invariant to size and color of the mouth. The mouth localization and tracking techniques are computationally efficient and the system recognizes the outer lip contours within a reasonable time.

7. REFERENCES

[1] Matthews, Iain, Timothy F. Cootes, J. Andrew Bangham, Stephen Cox, and Richard Harvey. "Extraction of visual features for lipreading." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, no. 2 (2002): 198-213.

[2] Zhi, Qi, A. D. Cheok, K. Sengupta, Zhang Jian, and Ko Chi Chung. "Analysis of lip geometric features for audio-visual speech recognition." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 34, no. 4 (2004): 564-570

[3] Siatras, Spyridon, Nikos Nikolaidis, Michail Krinidis, and Ioannis Pitas. "Visual lip activity detection and speaker detection using mouth region intensities." Circuits and Systems for Video Technology, IEEE Transactions on 19, no. 1 (2009): 133-137.

[4] Gao, Yongsheng, and Maylor KH Leung. "Face recognition using line edge map." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, no. 6 (2002): 764-779.

[5] Heisele, Bernd, Alessandro Verri, and Tomaso Poggio. "Learning and vision machines." Proceedings of the IEEE 90, no. 7 (2002): 1164-1177.

[6] Viola, Paul, and Michael Jones. "Robust real-time object detection." International Journal of Computer Vision 4 (2001).

[7] Wang, Lirong, Xiaoli Wang, and Jing Xu. "Lip detection and tracking using variance based Haar-like features and kalman filter." In Frontier of Computer Science and Technology (FCST), 2010 Fifth International Conference on, pp. 608-612. IEEE, 2010.

[8] Eveno, Nicolas, Alice Caplier, and P-Y. Coulon. "A parametric model for realistic lip segmentation." In Control, Automation, Robotics and Vision, 2002. ICARCV 2002. 7th International Conference on, vol. 3, pp. 1426-1431. IEEE, 2002.

[9] Yao WenJuan, Liang YaLing and Du Minghui. "A real-time lip localization and tacking for lip reading ." In Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on, vol. 6, pp363-366.

[10] Yong-hui, Huang, Pan Bao-chang, Liang Jian, and Fan Xiao-yan. "A new lip-automatic detection and location algorithm in lip-reading system." In Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on, pp. 2402-2405. IEEE, 2010.

[11] Zhang, Jian-Ming, Liang-Min Wang, De-Jiao Niu, and Yong-Zhao Zhan. "Research and implementation of a real time approach to lip detection in video sequences." In Machine Learning and Cybernetics, 2003 International Conference on, vol. 5, pp. 2795-2799. IEEE, 2003.

[12] Xu, Chenyang, and Jerry L. Prince. "Snakes, shapes, and gradient vector flow." Image Processing, IEEE Transactions on 7, no. 3 (1998): 359-369

[13] Liew, Alan Wee-Chung, Shu Hung Leung, and Wing Hong Lau. "Lip contour extraction from color images using a deformable model." Pattern Recognition 35, no. 12 (2002): 2949-2962.

[14] Caplier, Alice. "Lip detection and tracking." In Image Analysis and Processing, 2001. Proceedings. 11th International Conference on, pp. 8-13. IEEE, 2001.

[15] Gonzalez, Rafael Ceferino, and Richard E. Woods. Instructor's Manual for Digital Image Processing. Addison-Wesley, 1992.