Lip Detection and Tracking with Geometric Constraints under Uneven Illumination and Shadows

Waqqas ur Rehman Butt[1]
Higher Colleges of Technology Comp Info Sciences (CIS)
Ras Al Khaimah, UAE

Prof. Luca Lombardi[2]
Dipartimento di Ingegneria Industriale e dell'Informazione
University of Pavia, Pavia, Italy

***Abstract*—In the modern era, recent advancement in computer vision has led to emergent attention in lip reading. Indeed, lip-reading is used to understand speech without hearing it, and the process is mentioned as a lip-reading system. To construct an automatic lip-reading system, locating the lip and defining the lip region is essential, especially under different lighting conditions, significantly impacting the robustness of the lip-reading system. Unluckily, in previous studies, lip localization under illumination and shadow consideration has not been well solved. In this paper, we extant a local region-based approach towards the lip-reading system. It consists of four significant parts, firstly detecting/localizing the human face, mouth and lip region of interest in the first video frame. Secondly, apply pre-processing to overwhelmed the inference triggered by illumination effects, shadow and teeth appearance, thirdly create contour line using sixteen key points with geometric constraint and stored the coordinates of these constraints. Finally, track the coordinates of sixteen points in the following frames. The proposed method adapts to the lip movement and is robust in contrast to the appearance of teeth, shadows, and low contrast environment. Extensive experiments show encouraging results and the proposed method's effectiveness compared to the existing methods.**

***Keywords*—Lip detection; lip tracking; illumination equalization; shadow filtering; 16 points lip model**

I. Introduction

The continuous progress of technology brings to an irreversible change of paradigms of interaction between humans and machines. Traditional ways of human-computer interaction using keyboards, mice, and display monitors are being replaced by more natural modes, e.g. speech, touch, and gesture. 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 they will be completely transparent to users. In recent years, to automate the process of voice communication with which they interact between themselves persons. Lip movement and reading are used to recognize speech from a speaker without hearing. It is a procedure that especially gets to grips by people having hearing problems. In 1976, audio-visual illusion became recognize as the McGurk effect [1], which shows that visual cues information combined into the listener's mind automatically and unintentionally. The listener perceived the syllable, which is dependent on the visual information and strength of audio from the speaker.

In the past, there are two main techniques, edge and region-based, proposed for lip segmentation and extraction by using spatial information (edge and colour) to track lip movement. Hue and edge information are used to attain mouth localization and segmentation [2]. Initially, visual features extraction is obtained in greyscale images [3, 4]. The vertical center of the lip region is used to initiate by compelling the sum of each row in the mouth region and finding the minimum value of the row. The corners of the lips are found by setting the threshold, and horizontal edges are representing by four parabolas of both lips (lower and upper lips edges). They use the linear filter to find the edges. Another method is applied to the greyscale image [4], which is very close to the above method, but this approach tracks the unnecessary features in the mouth region such as nostril and pupils. The statistical colour model was used to locate the face by normalizing the skin colour [5]. Outliers are used to find the position of features points in all frames of the image sequence, and sometimes these positions are not the best. The performance of these techniques failed to produce an accurate result in cases when a speaker has beard and teeth presence. The beards have high edges in both directions (vertical and horizontal) mentioned in [6]. Therefore, the edge-finding method is not helpful for persons having bears. The HSI (hue, saturation and intensity) colour space extracts mouth pixels and sorts out the illumination from colours[7]. Hue values redefine the lip pixels. Different colour spaces [8, 9] and approaches have been used for visual feature extraction, e.g. optic flow analysis [10]. However, these methods failed with data sets of more than one few words and were computational intensive [11].

The active contour model (ACM) detects the lip boundary's edge [12]. Unluckily, this model often converges to the wrong result when the lip edges are indistinct, or the lip is very similar to the skin region. The region-based approaches mostly use the regional statistic characteristics to comprehend lip tracking. Distinctive examples include deformable template (DT) [13, 14], region-based ACM [15], active shape model (ASM) [16,17], and active appearance model (AAM) [18]. A regional cost function is used by DT to divider a lip image into the lip and non-lip regions via a parametric template, which represents the lip shape properly. Therefore, globally statistical characteristics have been used in mostly region-based approaches.

Consequently, their performance may decrease due to the appearance of teeth, tongue or black hole. The localized active contour model (LACM) [19] have better results. However, LACM depends on the proper correlative parameters.

Moreover, the colour information is not considered [19], which is very important to improve extraction performance, particularly when the images have shadows [20, 15].

This paper presented an approach to lip detection and tracking with two main phases: (i) lip contour extraction for the first frame and followed by (ii) lip tracking in the following lip frames. In the first phase, we created the dataset from a different speaker, i.e. Male/ female, different age groups by uttering English alphabets and numeric numbers in different light conditions, defining the mouth ROI, and applying pre-processing methods. Then, we utilized a 16-point lip model [21] with geometric constraints to achieve lip contour extraction. We repeat the same procedure for the lip ROI image and compute the lip tracking in the second phase. The proposed approach is adaptive to lip movement and robust against the appearance of the illumination effects, teeth and shadow. Experimental results have shown promising results.

II. Methodology

Previously, videos dataset created by using compression, controlled light environment, constant background for processing [22] led to noisy pixels in frame images, slow performance and caused false feature detection. We created the video datasets in different lighting conditions, gender (male/female) and different age groups. Some male persons had a moustache as well. A small application was developed for recording the video files by using Visual C++ and OpenCV [23]. These videos were recorded in the Computer Vision Lab of the University of Pavia at different times, using a Logitech HDR webcam with the highest possible resolution supported by the camera. Each speaker had to record the video by uttering different alphabet letters and numbers.

A. Face and Mouth Detection

Face and Mouth detection have a vital role in lip localization. Firstly, it is necessary to detect the speaker's face in all video frames and crop the speaker face for the mouth area. Numerous approaches have been already developed and categorized as: i) colour based [24], ii) template-based [25], and iii) feature-based [26,27]. Face detection methods based on local features and machine learning-based binary classification methods[28] have been widely used in various face recognition studies because of their real-time capability, high accuracy, and availability in the OpenCV, but the mouth area detection was not detected accurately. Only face detection results were accurate. To overcome this problem, we used the face image and split it into two parts horizontally. The upper part has the eyes, forehead and a small part of the nose. The lower part has the mouth on which we applied the mouth cascade classifier, obtaining precise results as shown in Fig. 1.

Fig. 1. Face and Mouth ROI Detection a)Frame Detection b) Face Detection c) Face Image Splitting d) Mouth Detection.

III. Pre-processing

A. Illumination Equalization

Mouth ROIs are extracted from videos acquired, where sometimes lightning is very strong and irregular. This irregularity is the cause of various disorders that can lead to malfunctions of the lip-reading application and make it challenging to identify the crucial points and construct the 16 points lip model. Different methods have been proposed for image enhancement [30], Histogram equalization, and lighting [20]. The method [29] works exclusively on the luminance value of the individual pixel. Although, It has few flaws, such as the effects of irregular lighting are attenuated only along with the single direction vertical and fixed scaling size of image 71 x 44 and mask size 3x3. We decided to improve the model [30], making it more robust with respect to light, multiple directions horizontal and vertical, working no more than on the single pixels and on local regions within the image. The extended algorithm can adapt to the multiple directions of the lighting, as shown in Fig. 2 [20].

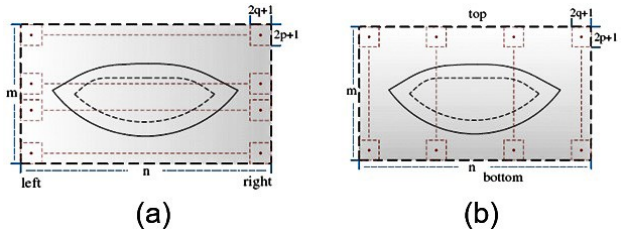

Fig. 2. a) Horizontal Direction b) Vertical Direction [20].

The colour lip image size m × n provided to the function's input is initially converted in HSV colour space, let L (i,j) is the L' (i,j) represents the luminance of each pixel respectively before and after the operation of equalization. To simplify the process, assume that the non-uniform illumination is instead linear along the direction of its application. As mentioned earlier, the innovation brought to a method implemented in this elaborate consists of manipulating the individual pixel's luminance value but work on a local region of size (2p + 1) × (2q + 1). Each pixel of the original image assumes the value obtained from calculating the average of the luminance values of all the pixels included in the mask that flows throughout the image along the two main directions identified. The luminance value of the pixels (Horizontal and Vertical directions) was calculated using the application formulated in formulas 1, 2.

$$\boldsymbol{L'}(i,j) = \begin{cases} L(i,j) + \frac{(n-2j+1)*(r(p)-l(p))}{2(n-1)}, & i \in [1,p] \\ L(i,j) + \frac{(n-2j+1)*(r(m-p)-l(m-p))}{2(n-1)}, & i \in [p,m-p] \\ L(i,j) + \frac{(n-2j+1)*(r(i)-l(i))}{2(n-1)}, & i \in [m-p,m] \end{cases} \quad (1)$$

$$\boldsymbol{L'}(\boldsymbol{i},\boldsymbol{j}) = \begin{cases} L(i,j) + \frac{(m-2i+1)*(b(q)-t(q))}{2(m-1)}, & i \in [1,q] \\ L(i,j) + \frac{(m-2i+1)*(b(j)-t(j))}{2(m-1))}, & i \in [q,n-q] \\ L(i,j) + \frac{(m-2i+1)*(b(n-q)-t(n-q))}{2(m-1)}, & i \in [n-q,n] \end{cases} \quad (2)$$

Where l_i and r_i denote the average intensity of respectively left and right edges of the local region of size (2p + 1) × (2q + 1), to the i_{th} row of the mask. Similarly, t_j and b_j denote the average intensity of the upper and lower edges of the local region at the j_{th} column of the mask, as shown in Fig. 7.

B. Teeth Filtering

During the experiment, it was observed that illumination equalization is not enough to improve the system. Still, some other factors are to be considered, e.g. the teeth, black hole, and tongue region that can be visible in processed images. In the past, the researchers reported that without considering these factors cannot have a robust result for lip tracking [31]. The proposed teeth filtering method removes the teeth appearing in the mouth (ROI) in all frames. The presence of teeth in the image frame is observed when the mouth status is open. It was possible to construct a filter dependently on thresholds, which correctly identifies the range of colour that characterizes the range of the tooth region in the Mouth (ROI). The implemented function inputs the illumination equalized image and then convert it into two different colour spaces *CIELAB* and *CIELUV*. *RGB* colour space has characteristics that are not suitable for defining the thresholds based on which it filters the region of teeth. The teeth region is characterized by the lowest components a∗ and u∗ present in the image. The two threshold values have been set for the two chromatic components of interest. L_{au} has demonstrated that to achieve a satisfactory result, the teeth thresholds t_a and t_u should be set according to the formulas (3) [41].

$$t_a = \min(\mu_a - \sigma_a, 9] \text{ Otherwise if } (\mu_a - \sigma_a) < 9$$

$$t_u = \min(t_u - t_u, 29] \text{ Otherwise if } (\mu_u - \sigma_u) < 29 \quad (3)$$

Where μ_a, σ_a, μ_u and σ_u are, respectively, the mean and standard deviation of the chromatic components to a∗ and u∗. According to this approach, all the pixels that relate to the teeth may identify pixels that correspond to the teeth i.e. $a^* < t_a$ or $u^* < t_u$ or L^* 35% or L^* 95% in the chromatic components reference as white, normalized to restrict their range of the standard deviation in 2 around the mean value. Each pixel characterizing the teeth is masked by resetting the colour value of all the chromatic components of the specific colour space identified. It cannot influence the future operations of search from the position of the mouth. If there are no teeth, a presence mask will not apply, and if teeth appear, the mask will remove the teeth pixels by changing the teeth pixels value to 0.

C. Shadow Filtering

A mean filter is used to reduce noise caused by shadows in the images described [32]. Again, these disturbances are because of different lighting conditions, e.g. sourcing light angles and shades beneath the lower lips. The rate of recognition of a lip-reading system is based on the accuracy of the lip position. Unfortunately, up to now, there are no algorithms not effectively solved the problem of locating the lips in uneven lighting conditions. Illumination equalized images used to reduce the interference brought by shadow. For the implementation of the shadow detection method, we used these steps: i) Convert the illumination equalized image into grayscale, ii) Considering the image as a matrix in which the rows are characterized by the index "i" and columns from the index j. iii) Calculate the accumulation of the grey level value for each column of the image and obtain a column index corresponding to the mean value of the accumulation curve as the boundary of shadow. It is used to divide the grayscale image into two sub-images, left Isl and right Isr, to enhance the contrast between lips and the surrounding skin region (4).

$$I_e = \frac{255(I - I_{min})}{(I_{max} - I_{min})} \quad (4)$$

$$\delta_i = dist(I^{(i+1)}), \quad (I^{(i)}) \quad (5)$$

Where, I_{min} is the minimum grey-level value in the image, and I_{max} is the maximum value. Euclidean distance δ_i is determined by calculating the distance between (I^t) and (I^{i+1}). If δ_{i+1} is greater or equal to δi then the process will stop, and (Ii) will be marked as the final image. The convolution process ends when the Euclidean distance decreases by less than two units between two subsequent iterations. We determined if each sub-image and the whole image and output image Isl are extracted by subtracting the initial and final images in the proposed function. The shadow detection ends by making a new image by merging two images, left Isl and right Isr. The middle line obtained between Isl and Isr by curve, having information about boundaries of lips and skin region and the minimum value of the row position, is considered the corner points of the lips. Finally, a convoluted image is extracted, as shown in Fig. 3.

Fig. 3. Smoothing the Contrast (a) Greyscale Image (b) Left Image (I_{sl}) Convoluted, (c) Right Image (I_{sr}) Convoluted (d) Output Image.

IV. Lip Detection

In this step, we have to mark the exact position of the lips. An elliptic shape function [36] was applied to detect the lip boundary. This method gives good results when the mouth status is closed, but when the variation in lips, some marginal parts of this elliptic region may be far away from the lip boundary. Lip corner dots are successfully implemented by using intensity variation and colour cues in [19, 33], as shown in Fig. 4(a). We proposed the extraction method for geometric positions by labelling the left corner, right corner, upper corner and lower corner as points L_a, L_b, V_a and V_b as shown in Fig. 4(b).

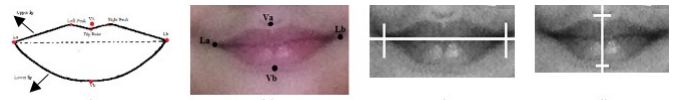

Fig. 4. a) Standard Lip Model b), Geometric Points of Interests, c) Crucial Points Horizontal d) Crucial Points Vertical.

A. Crucial Points (L_a, L_b, V_a, V_b)

To identify the horizontal crucial points; we extracted the left point L_a and right point L_b. Points are located on the median axis of the image and stored in a vector. The values of curve G have peaks of high frequency due to noise. This noise must be clear to ensure the precise extraction process of crucial

points. A low-pass filter applied of Butterworth through mask size 3 3 pixels that run through everything the curve G to reduce the noise. The vector filtered result, called Gf proceed with the search of the crucial points. Accumulate the grey-level value for each column of the image and obtain correspondence column index to the mean value of the accumulation curve as the boundary of shadow. The boundary shadow value, median axis and sub-images Isl and Isr already have from the shadow filtering phase. This whole curve G consists of both sub-images divided by boundary shadows. The curve representing the vector G should be monotonically decreasing to quickly identify the crucial point left as that point occurs the maximum value of the gradient, considered absolute value. This value corresponds to the boundary between the mouth region and skin. The positions of mouth corners correspond to the steep slopes of the curve, as shown in Fig. 4(c,d). The method for searching for the left crucial point consists of several steps. Firstly, obtain the first minimum' m' by scanning Gf from left to right. To work in the best possible conditions, make the monotonic curve and save this new curve in a vector support Gm using formula (6).

$$G_m^{(i)} = \begin{cases} G_f^{(i)} & (G_f^{(i)} \geq G_f^{(i+1)}) \\ G_f^{(i+1)} & (G_f^{(i)} < G_f^{(i+1)}) \end{cases} \quad (6)$$

Once curve is extracted, carry out all the values of the vector Gm in the correct range of processing for grayscale images, therefore ensuring the pixels fall in the range between 0 and 255 by using the following formula (7).

$$G_m^{(i)} = \frac{255(G_m - G_m^{(min)})}{G_m^{(max)} - G_m^{(min)}} \quad (7)$$

Gm is used to search the left crucial point, but unfortunately, there is no maximum rate, although the curve is monotonic. To overcome this problem, calculate the average pixels values of the vector Gm as shown in equations (8) and (9) are utilized to adjust the image's contrast on the horizontal median axis.

$$C^{avg} = \frac{\sum_{i=1}^{k} C'}{K} \quad (8)$$

$$I_{out} = \begin{cases} 255 & 1.5C^{avg} < I_{in} < 1 \\ \\ \frac{500}{c^{avg}}, & 0 < I_{in} < 1.5\, C^{avg} \end{cases} \quad (9)$$

Where I_{out} and I_{in} are the output and input grey level values, the I_{out} is obtained by adjusting the curve and a binary image; those pixels values are 0 or 255. The local minimum point is identified in the position for the first time, where the pixel changes from 255 to 0. Curve C was obtained after the adjustment of contrast and crucial points (L_a , (L_b). The vertical points made a start from the results as described in horizontal crucial points. Based on the position of the horizontal crucial points, calculate the mouth's centre point, and its column index is marked as vertical midline of the mouth, in which two crucial points, vertical Va and Vb are situated. The pixels values that lie on the vertical axis are divided into two groups and stored on a vector to be processed. We built two more vectors, respectively called B1 and B2, containing only the pixels with a value equal to the maximum, corresponding to the value "1", and the pixels with a value equal to the minimum conform to the value "0". In addition, two binary vectors B_1' and B_2' are obtained by applying logical operation that provides outbound B1 and B2 as described in equation (10).

$$B_1' = B_1\, XOR\, (B_1 \ll 1) \quad B_2' = B_2\, XOR\, (B_2 \ll 1) \quad (10)$$

The operator $\ll$ indicates the logical operation of shift one position to the left. At this stage, crucial points Va and Vb are identified. These points where the first occurrence of the value of $'1'$ inside of the vector B1′ while Vb fits in the position in which, the last occurrence of the value $'1'$ inside of the B2′.

B. Draw Ellipse

The next step is to find an ellipse that encloses the mouth region. In some cases, the ellipse position is incorrect because the bottom point V_a is not proper. To overwhelm this problem, draw two half-ellipses, one for the upper lip and one for the lower lip. This trick shows more precise and realistic results. This method identifies the coordinates xI, yI of the horizontal crucial points Lax, Lbx and Lay, Lby and vertical crucial points Vax, Vbx and Vay and Vby by using geometrical formulas for drawing an ellipse. The centre of the mouth is calculated with equations (11).

$$x_c = \frac{1}{2}\,(La_x + Lb_x\,) \quad y_c = \frac{1}{2}\,(La_y + Lb_y\,) \quad (11)$$

The inclination of the half-ellipses for the horizontal plane is calculated with equation (12).

$$\theta = \arctan\left(\frac{La_x + Lb_x}{La_y + Lb_y}\right) \quad (12)$$

The semi-major axis a common to both the half- ellipses, calculated as in formula (13). The semi-ellipse of the upper lip and the lower lip's semi-ellipse is shownin equation (14).

$$a = \frac{1}{2}\,\left(Lb_x - La_x\right)^2 + (Lb_y - La_y)^2\Big)^{1/2} \quad (13)$$

$$b_{up} = \frac{1}{2}\,\left(Va_x - x_c\right)^2 + (Va_y - y_c)^2\Big)^{1/2}$$

$$b_{low} = \frac{1}{2}\,\left(Vb_x - x_c\right)^2 + (Vb_y - y_c)^2\Big)^{1/2} \quad (14)$$

xc,yc be the centre of mouth coordinates and origin of the combined semi-ellipse, calculated as formula (15). Where 'a' is the semi-major axes, bup and blow are the upper and lower semi-minor axes. θ is the inclined angle, defined at the counter-clockwise direction.

$$\frac{x_{up}^2}{a^2} + \frac{y_{up}^2}{b_{up}^2} = 1 \qquad \frac{x_{low}^2}{a^2} + \frac{y_{low}^2}{b_{low}^2} = 1 \quad (15)$$

C. Lip Modeling

The Lip model was used to determine the accurate boundary line and geometric points around the lips. We have already extracted four points left, right, top and bottom in the previous section. Previously, lip modelling was performed without pre-processing, which may cause incorrect tracking results, i.e. four key points model with two parabolas for the lip

contour used in [34] and six key points with cubic curves connected to describe the lip shape used in [35]. First, sixteen point geometrical deformable models are used in [21]. It is challenging for the modelling of the lips in non-ideal conditions. A model-based approach was proposed for lip contour extraction from colour images to overcome this problem. A region-based cost function is employed to formulate the entire lip contour extraction as a region partition problem instead of the conventional edge detection problem. The proposed algorithm is more robust with low colour contrast, and the final extraction result is less sensitive to the initial model parameters than the edge-based. Curve C0 is used as a curve of evolution for the initial model development of the model. The proposed algorithm is the extension of the 16 points lip model as described in [21]. These geometric constraints showed the lip boundary, and we will store the location of these points and then track the lip movement. The 16 lip boundary points labelled P0 to P15 in anti-clockwise and parameter set by equation (16).

$$\lambda_p = \{x_{pi}, y_{pi}\} \quad where\ i = 0, \ldots .15 \tag{16}$$

These points are divided into three groups as the lower lip (P_0,P_7,P_{15}), upper right lip (P_7- P_{11}) and upper left lip (P_{11}- P_{15}). A normalization process was used to translate the lip corner points P_7 and P_{15} to lie on the horizontal x-axis and point P_{11} on the vertical y-axis. The centre origin of lips is set to be the midpoint between the two lip corners, P7 and P15. After normalization of the mouth ROI, the next step is constructing the 16 points lip-model lips. Lip modelling is split into two parts model i) initialization and ii) thresholding.

- Lip Model Initialization

Elliptical regions extract the lip contours [36] but give the approximate surrounded area of lips, not precise lip. Therefore, a minimum-bounding ellipse as the initial evolving curve is used to find the extract of the lip contours. Model initialization is the starting point of the construction of the lip model. Using the ellipse's geometric parameters is already identified, and the ellipse showed the accuracy of locating the mouth region in the video frames. Therefore, some adjustment operations are required to simplifying the process of initialization of the model. We used three functions, probability map, cost function formulation and draw graph to obtain more accurate model construction.

The teeth pixels with a∗ < ta or u∗ < tu for colour component ‘a’ and ‘u’, also white pixels of L∗ ≤35 marked and discarded in lip initialization process and assign probability values on marked pixels. Teeth pixels are always inside the mouth and assigned high values of probability. It helps the model to separate the upper and lower lip. The surrounding teeth are considered lip pixels with low luminance values and equal probability values 0.5 are assigned. A low pass filter and cost function are applied to smooth the probability map to optimise the process and determine the optimum partition when the cost function in equation (17) is maximized [21].

$$max\ \Big\{C(\lambda_p) = \prod_{(x,y)\epsilon\, R_l(\lambda_p)} prob_l(x,y) * \prod_{(x,y)\epsilon\, R_{nl}(\lambda_p)} prob_{nl}(x,y)\Big\} \tag{17}$$

Where λp is the 16 point model parameters, probl (x, y) and probnl (x, y) are the probabilities of lip pixels and non-lip pixels at location (x, y), Rl and Rnl are the enclosed and outside the region by the point model. The lip model fitness evaluated by extension of the cost function. Draw Graph proposed to draw the graph extraction points starting from the parameters of the lips. Draw the rectangle around the mouth, tolerance value calculated based on the image's height, and enlarge a variable value the rectangle within which to seek the edges of lips. The rectangle is determined by the ellipse, which allows the search of the representative point of the intersection with the green ellipse. Calculate the center point of La, Lb and Va, Vb crucial points and mark the points as shown in Fig. 5. The model initialization worked with two single-channel images ’H’ of HSV obtained from the original RGB and Ellipse image. The final mouth ROI image is obtained by subtracting the ellipse H image from the original input H image.

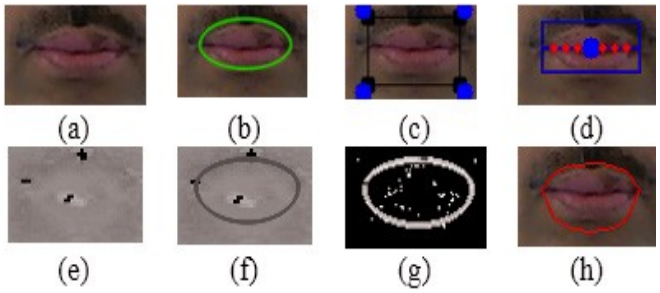

Fig. 5. a) Original Image, b) Ellipse Image, c) Draw Rectangular d) Graph Points, e) ’H’ Channel Image (b), f) ’H’ Channel Image of Ellipse *’RGB’* Image , g) Subtracted Image, h) Final Result.

- Lip Model Thresholding

In this section, we refine the position of the lips points and store the coordinates (x, y) of all points. These coordinates are stored in a file and later will be used in the lip-tracking phase. The channel ‘H’ is used to search the points of the model, and this process is implemented similarly to model initialization. The segmentation was carried out as follows: the red colour of channel H was exploited to find better positions for each point. We looked for neighbour pixels of each point; if there was a significant variation of the red colour (lips to the skin), upgrade the position of the point. The tolerance value for the lower and upper lips have already been calculated. The points P15, P7 have already been found and used in the same position. The position of the remaining 14 points will be upgraded by using means of this procedure. We divided the image into four parts. The lower lip boundary with white points and the upper lip boundary with yellow points are clearly visible. We added the two horizontal corner points P15,P7. Finally, the resulting image has 16 points, as shown in Fig. 6(e,f).

D. Lip Tracking

We have been extracted the 16 points with their positions and coordinates of the first frame. And then tracking algorithm is applied to track the movements of the lips with these points in the subsequent frames. They assumed that frames are extracted from the same video sequence and have almost the same geometric characteristics. It was decided to simplify the extraction phase of the contour of the lips in the remaining

frames to considerably overcome the computational algorithm's burden. Pre-processing is carried on all frames. The bulk of the computational algorithm has segmented the image to identify the mouth. Once this process is completed for the first frame, the differences between two consecutive frames are minimal and are limited exclusively, e.g. the lips assume during speech. In this way, it is limited processing on the second frame to a portion of the image significantly smaller, as shown in Fig. 6(e), (f). The construction of the 16 points lip model on the second frame in the sequence is applied exclusively in the image identified. We have coordinates of lips points in the first and second frames. Table I is showing the (x, y) coordinates of two frames. The first frame is the initial frame when the mouth status closes, and the second frame when the speaker said alphabet B, the mouth status is changed and has different coordinates. The upper part of the lips is moved because slight variation is obtained in P5, P4, P6, P15 points, and considerable variation in P9, P13, P14 points.

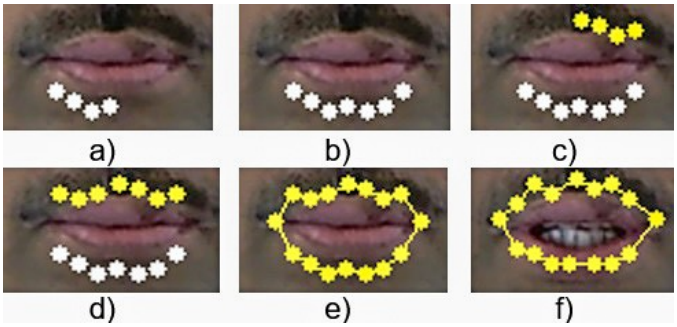

Fig. 6. Drawn Points of the Four Parts of the Image (a)4 Left Lower Points (b) 3 Right Lower Points (c) 4 Right Upper Points (d) 3 Left Upper Points (e) Final Result 1st Frame f) Final Result 2nd Frame.

TABLE I. COORDINATES OF 16 POINTS FIRST AND SECOND FRAME

First Frame			Second Frame		
Point	**X**	**Y**	**Point**	**X**	**Y**
0	21	34	0	18	32
1	28	38	1	25	34
2	35	42	2	32	38
3	42	40	3	39	38
4	50	42	4	49	38
5	57	40	5	56	38
6	64	34	6	63	34
8	64	10	8	63	12
9	57	12	9	56	6
10	50	8	10	49	8
11	43	6	11	42	4
12	35	10	12	32	10
13	28	12	13	25	6
14	21	10	14	18	14
7	72	21	7	74	20
15	14	21	15	11	20

V. RESULTS AND DISCUSSION

All the frames were processed in two sizes 150x150 and 360x360. The results showed that it was possible to extract the contour of the lips precisely in many of the data sets, but some experiments did not go well, where the software did not work correctly (about 5% of cases). The lip tracking procedure is applied to sequences of frames, starting with the first frame to complete the entire sequence. The average processing time for performing lip tracking is about 0.556 seconds to extract the contour of the lips from the first image of the sequence and approximately 0.09 seconds for subsequent frames. Such times include the operations of reading and writing the images used to test the algorithm. It is estimated that the video stream can achieve a frame rate of about 12-15 fps (frames per second) to allow a real-time execution. Face detection in videos was our first step to build the speech recognition system. Secondly, the mouth was detected in the face image, and mouth ROIs were defined for further processing. The accuracy of the classifiers is described in Section II(A), as shown in Table I. To increase the accuracy rate for mouth detection, we improved the mouth cascade and compared it with previous classifier results, as listed in Table II. The proposed method for mouth detection showed more precise results as compared to the earlier methods.

Furthermore, an effective method to reduce the effects of uneven illumination is proposed as more robust to light, multiple directions (horizontal and vertical), and working on both the single pixels and on local regions within the image. Fig. 7 is showing the effectiveness of the proposed illumination equalization method. There are some darker parts on the left or right side, and the same in vertical directions where darker parts are on the top or bottom. The darker part was significantly reduced by applying the proposed method, as shown in Fig. 7(d).

In the previous lip-reading system, teeth masking was not considered, as it was found based on local regions and exploited information on colour inside the mouth. The proposed teeth detection method successfully removes the teeth area in the mouth ROI in all frames and pixels belonging to teeth removed, as shown in Fig. 8(a).

TABLE II. COMPARISON OF MOUTH DETECTION

Facial Features	Positive Hit Rate %	Negative Hit Rate %
Face	90	29
Mouth [29]	67	28
Proposed	86	19

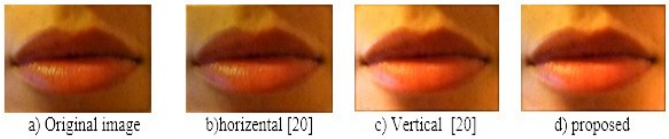

Fig. 7. Comparison of Illumination Equalization.

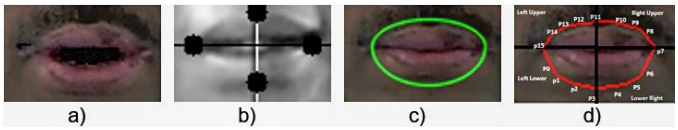

Fig. 8. Proposed Method Results a) Teeth Filtering b)Crucial Points c) Semi Ellipse d) 16 Points Lip Model.

In the past, shadow filtering methods have been proposed for indoor environments and have not been studied for lip localization. Therefore, we proposed a technique for shadow removal as described in the shadow filtering section. After applying the pre-processing steps, a filtered image was obtained where uneven equalization, teeth filtering, and shadows were significantly reduced. In the next step, mark the exact position of lips, which leads to lip detection and tracking. An elliptic shape function is used for lip detection [12]. In this method, lip detection is correctly performed when the mouth is closed; on the other side, some parts of the lip region are discarded during lip movement. Lip corner dots are successfully implemented by using intensity variation and colour cues in [37, 38]. These dots do not fit the exact geometric position of lip structure. Therefore, We proposed a method for extracting the horizontal and vertical positions of median axes of the mouth by labelling the left, right, upper and lower corners as points La, Lb and Va ,Vb. The results of crucial points as shown in Fig. 8(b). Crucial points are needed to draw the ellipse. The upper lip has three corners: left, right, and dip points due to Cupidon's bow. Fig. 8(c) is showing the result of the proposed combined semi ellipses methods. Different lips models were applied, e.g. four key point models and six key point models [34,35]. We used the 16 points, lip model. The elliptical regions extract the lip contours [12], but they do not give precise lip contours. Therefore, the combined ellipse is used as the initial evolving curve to find lip contours in the proposed method. Some adjustment operations are applied to simplifying initialization by using probability map, cost function formulation and draw a graph to obtain more accurate model construction. A Thresholding function was used to refine the initial evolving curve image better to get a precise position. The lip image is divided into four parts: Left lower P0-P3, Lower right P4-P6, Right upper P8-P10 and Left upper P11-P14 and stores the coordinates of each point, as shown in Fig. 8(d).

Table III described the computation time obtained by the proposed method for the first frame as 0.556 seconds, which is smaller than the methods' results described in [21, 35, 15]. The average computation time of lip tracking for one frame is 0.099 seconds, which is less than the methods' values [21, 35, 15]. The method [21] needs to compute the probability map at every frame. The method [35] requires a bit more pre and post-processing techniques and adjustment processes to fit the lip boundary. However, the average computation time of tracking one lip frame is higher than [38].

The average extraction performance and lip boundary extraction degraded due to low contrast, uneven lighting conditions and irregular shapes in the lip image. The proposed pre-processing methods are applied to the complex appearance of shadows, uneven illumination and teeth. These factors are considered, and the extraction performance of lip boundary extended up to 96%, as shown in Table IV.

TABLE III. COMPARISON OF COMPUTATION TIMES (SECONDS)

Frames	**Barnard et al. [38]**	**Wang et al. [21]**	**Eveno et al. [35]**	**Yiu et al. [15]**	**Proposed**
First	Manual	1.232	0.623	0.695	0.556
Tracking	0.0989	0.133	0.171	0.103	0.099

TABLE IV. COMPARISON OF FEATURES EXTRACTION PERFORMANCE (%)

Data set	Kass et al.[12]	Liew et al. [20]	Leung et al.[39]	Werda et al. [40]	Proposed
	79.50	92.50	89.00	91.00	96
Average	76.7	89.57	83.71	88.57	96

VI. CONCLUSION

An approach to detect and track lip boundaries is presented that highlights the lips and avoids other factors, e.g. false lip pixels and recovers from failures. The proposed algorithm comprised the lip-tracking module from the lip boundary lines, a feature vector of 16 points lip model. Three pre-processing steps, illumination equalization, teeth detection, and shadow removal, aim to investigate edge information and global statistical characteristics. The lip tracking method used 16 points lip model on the lips, stores the coordinates of these points and tracks these coordinates during the utterance by the speaker. Moreover, the proposed method is easy to implement and computationally efficient, capable of locating face and mouth and lips feature points that give a high accuracy rate for lip localization, modelling and tracking accuracy.

Experiments have shown that outliers detecting and better predicting ROIs can reduce the number of frames with locating or tracking failures. This research work brings together new methods, representations, and insights, which are quite generic and may have broader applications in computer vision, image processing, and speech recognition.

REFERENCES

[1] McGurk, H and MacDonald, J, "Hearing lips and seeing voices," Nature, vol. 264, pp.746-748, 1976.

[2] X. Zhang, R.M. Mersereau, "Lip feature extraction towards an automatic speechreading system," Proceedings of the IEEE International Conference on Image Processing, vol. 3, pp.226-229, 2000.

[3] R. Rao and R. Mersereau, "Lip modelling for visual speech recognition," In 28th Annual Asimolar Conference on Signals, Systems, and Computers, IEEE Computer Society, Pacific Grove CA, vol. 2, 1994.

[4] R. Steifelhagen, J. Yang and U. Meier, "Real time lip tracking for lipreading," In Proceedings of Eurospeech, 1997.

[5] J. Yang and A. Waibel, "A real-time face tracker," InProc. WACV, pp.142-147,1996.

[6] T.W. Lewis and D.M.W. Powers, "Lip feature extraction using red exclusion," ACM International Conference Proceeding Series, Sydney, Australia, vol. 9, pp.61-67,2000.

[7] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol. 3(1), pp.71-86,1991.

[8] Luca Lombardi, Waqqas ur Rehman Butt, Marco Grecuccio, "Lip Tracking Towards an Automatic Lip Reading Approach," Journal of

Multimedia Processing and Technologies, vol. 5, pp.1-11. ISSN: 0976-4127, 2014.

[9] 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, vol. 5, pp.2188-2196,2013.

[10] W.U.R. Butt, L. Lombardi, "A Survey of Automatic Lip Reading Approaches," 8th ICDIM 2013 (The 8th International Conference on Digital Information Management) in Islamabad, Pakistan, pp.299-302,2013.

[11] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," Int. J. of Computer Vision, vol. 12(1), pp.43-77,1994.

[12] M. Kass, A. Witkin, D. Terzopoulos , "Snakes: active contour models," International Journal of Computer Vision, vol. 1(4), pp.321-331,1988.

[13] Freedman, M.S. Brandstein, "Contour tracking in clutter: a subset approach," International Journal of Computer Vision, vol. 38(2), pp.173-186, 2000.

[14] A.L. Yuille, P.W. Hallinan, D.S. Cohen, "Feature extraction from faces using deformable templates," International Journal of Computer Vision, vol. 8(2), pp.99-111,1992.

[15] Yiu-ming Cheung, Xin Liu, Xinge You, "A local region based approach to lip tracking," Pattern Recognition, vol. 45, pp.3336-3347,2012.

[16] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active Shape Models ASM their training and applications," Computer Vision and Image Understanding, vol. 61(1), pp.38-59, 1995.

[17] J. Luettin, N.A. Thacker, and S.W. Beet, "Active Shape Models for Visual Speech Feature Extraction," .G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, NATO ASI Series, Berlin, Germany, vol. 150 , pp.383-390,1996.

[18] I. Matthews, T.F. Cootes, S. Cox, R. Harvey, and J.A. Bangham, "Lip reading using Shape, Shading and Scale," In D. Burnham, J. Robert-Ribes, and E. Vatikiotis- Bateson, editors, Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98, pp.73- 78,1998.

[19] X. Liu, Y.M. Cheung, M. Li, H. Liu,"A lip contour extraction method using localized active contour model with automatic parameter selection," Proc. of the IEEE International Conference on Pattern Recognition, pp.4332-4335,2010.

[20] A.W.C. Liew, S.H. Leung, W.H. Lau, "Lip contour extraction from colour images using a deformable model," Pattern Recognition, vol.35 (12). pp.2949-2962,2002.

[21] S. Wang, W. Lau, S. Leung, "Automatic lip contour extraction from colour images," Pattern Recognition, vol.37(12). pp.2375-2387,2004.

[22] Ahmad B. A. Hassanat, "Visual Speech Recognition," Speech and Language Technologies, Prof. Ivo Ipsic (Ed.), InTech, ISBN: 978-953-307- 322-4,2011.

[23] Opencv Library, "Open Computer Vision Library," http://sourceforge.net/projects/.

[24] R.-L. Hsu, Mohamed Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24(5). pp.696-706,2002.

[25] I. Craw, D. Tock, and A. Bennett, "Finding face features," In Proc. European Conference on Computer Vision, pp.92-96,1992.

[26] B. Heisele, T. Serre, M. Pontil, and T. Poggio, "Component-based face detection," In Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol.1, pp.657-662,2001.

[27] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: the importance of good features," In Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol.2, pp.53-60,2002.

[28] P. A. Viola and M. J. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol.57(2), pp.137-154,2004.

[29] Cristinacce, D. and Cootes, T, "Facial feature detection using AdaBoost with shape constraints," British Machine Vision Conference, 2003.

[30] Manvi, Rajdeep Singh Chauhan, Manpreet Singh, "Image Contrast Enhancement Using Histogram Equalization ," International Journal of Computing Business Research, pp.2229-6166,2012.

[31] Permit, P., "An Automated Visual Speech Reading System," PhD Thesis, Department of Computing, Communications Technology and Mathematics London Metropolitan University,2007.

[32] D. Xu, J. Liu, X. Li, Z. Liu, X. Tang, "Insignificant shadow detection for video segmentation," IEEE Transactions on Circuits and Systems for Video Technology, vol.15, pp.1058-1064,2005.

[33] M. Li, Y.M. Cheung, "Automatic Lip localization under face illumination with shadow consideration," Signal Processing , vol.89(12), pp.2425-2434,2009.

[34] Y. Tian, T. Kanade, J. Cohn, "Automatic Robust lip tracking by combining shape color and motion," Proceedings of the Asian Conference on Computer Vision, pp.1040-1045,2000.

[35] N. Eveno, A. Caplier, P.Y. Coulon, "Accurate and quasi-automatic lip tracking," IEEE Transactions on Circuits and Systems for Video Technology, vol.14(5), pp.706-715,2004.

[36] E.D. Petajan (1984), "Automatic Lipreading to Enhance Speech Recognition," PhD thesis, University of Illinois at Urbana-Champaign,1984.

[37] S. Lucey, S. Sridharan and V. Chandran, "Chromatic lip tracking using a connectivity based fuzzy thresholding technique," In ?ISSPA?99, vol.2, pp.669-672,1999.

[38] B. Mark, H. Eun Jung, O. Robyn , "Lip tracking using pattern matching snakes," Proceedings of the Asian Conference on Computer Vision, 2002, pp.23-25,2002.

[39] S.H. Leung, S.L. Wang, W.H. Lau, "Lip image segmentation using fuzzy clustering incorporating an elliptic shape function," IEEE Transactions on Image Processing, vol.13(1), pp.51- 62,2004.

[40] S. Werda, W. Mahdi, A. Ben Hamadou, "Colour and geometric based model for lip localisation: application for lip-reading system," Proceedings of the International Conference on Image Analysis and Processing, pp.9-14,2007.

[41] A. W. C. Liew, Shu Hung Leung and Wing Hong Lau, "Segmentation of color lip images by spatial fuzzy clustering," IEEE Transactions on Fuzzy Systems,vol.11, pp.542-549,2003.