New Lips Detection and Tracking System

Siew Wen Chin*, Kah Phooi Seng, Li-Minn Ang, King Hann Lim

***Abstract*— An automatic lips detection and tracking system based on watershed segmentation and H_∞ approach is presented. For some of the lips detection systems, skin/non-skin detection is a prerequisite step to localise the face region and the lips region is then detected from the face region. In this paper, a direct lips detection technique using watershed segmentation without needing preliminary face localisation is proposed. The watershed algorithm segments the input image into regions. Consequently, the cubic spline interpolant lips colour modelling and symmetry detection are used to detect the lips region from the segmented regions. The position of the segmented lips is passed to the H_∞ tracking system to predict the location of the lips in the succeeding video frame. The simulation results have revealed a good performance of the proposed method.**

***Index Terms*— Audio-visual speech recognition, H_∞ filter, Lips detection and tracking, Symmetry detection, Watershed.**

I. INTRODUCTION

In recent years, problems in the automatic speech recognition (ASR) have cropped up and drawn the attention of researchers [1]-[3]. With the presence of noise as in real world circumstances, the ASR rate could be dramatically reduced. The ASR system would be able to provide an appreciable performance only under a certain controlled environment. With the inspiration of lips-reading capability from the impaired society and the limitation of the noise robust techniques, the audio-visual speech recognition (AVSR) has become a research trend and is growing rapidly [4].

Dealing with the aforementioned AVSR, the front end lips detection and tracking is a key to make the overall AVSR system a success. J.M Zhang *et al.* [5] proposed a lips detection technique using red exclusion and Fisher transform. The skin-colour model and motion correlation are first applied in the system to locate the face region. The lips image is then enhanced by applying the red exclusion method, and the lips region is finally separated from the skin image via further thresholding. Jamal *et al.* [6] presented lips detection in the normalised RGB colour scheme. The incoming colour image is first normalised using either the pixel or maximum intensity normalisation scheme. The normalised image is then segmented into skin and non-skin regions using histogram thresholding and the lips detection is performed on the skin pixels. A tracking and lips feature extraction algorithm for speaker identification is proposed by T. Wark *et al* [7]. From the chromatic facial image, the region of interest, which is the lips region, is obtained and the contour model is then derived from the syntactic information. The speaker dependent feature vectors are then extracted from the colour information in and around the lips.

For the lips detection algorithms presented in [5]-[6], a preliminary face localisation is required for the lips detection and segmentation process. In this paper, instead of having the aforementioned face localisation process at the early stage, a direct lips detection and segmentation system based on watershed scheme is proposed. The overview system flow of the proposed system is depicted in Fig. 1. The initial video frame is first sent for pre-processing which includes: edge detection, obtaining foreground object and background ridge line. Subsequently, the outputs from both processes are passed to the watershed segmentation system [8]-[9]. From the resultant segmented regions, the lips region is detected by applying cubic spline interpolant lips colour modelling. If more than one region is detected, a further lips verification process is applied by using symmetry detection.

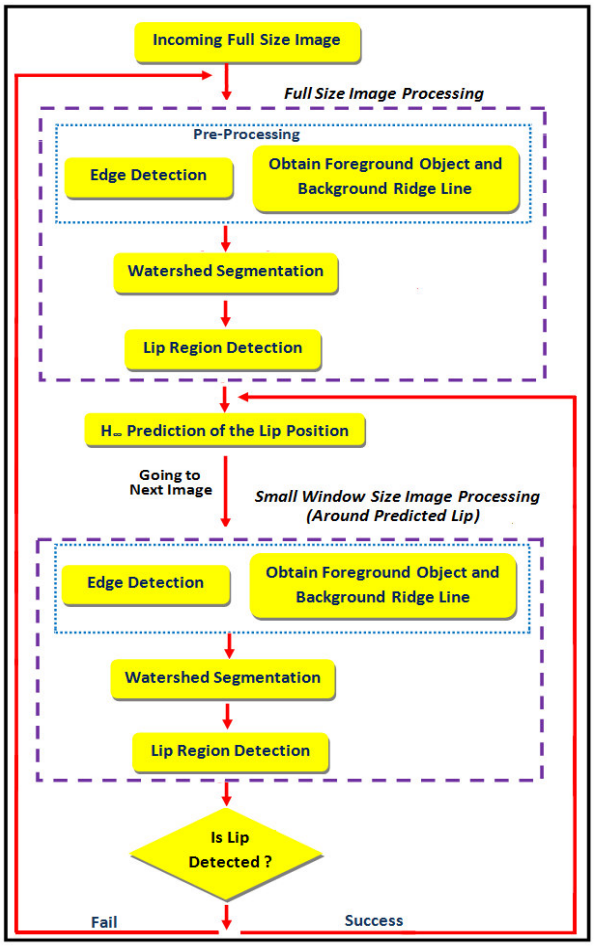

Fig. 1. Overall system flow of the lips detection and tracking system

Manuscript received October 13, 2008. Manuscript submitted to the IAENG International Conference on Image Engineering, ICIE'09.

The authors are with the School of Electrical and Electronics Engineering, The University of Nottingham, 43500 Semenyih, Selangor, Malaysia. Tel: +603 8924 8350, Fax: +603 8924 8071, Email: (keyx8csw, Jasmine.Seng, Kenneth.Ang, keyx7khl)@nottingham.edu.my

The centre point of the successfully detected lips region is forwarded to the H_∞ tracking system. By using the existing lips location, H_∞ filter is applied as the estimator to estimate the lips location in the subsequent frame. For the successive frame, the watershed segmentation and lips detection is only applied to the small window image size around the predicted location. If the lips position is predicted wrongly, full image processing would restart again.

II. The Proposed Lips Segmentation And Detection System

For the watershed algorithm, a grayscale image would be treated as a topographic surface. Every pixel is situated at certain altitude as a function of its gray level; black is regarded as minimum altitude while white as the maximum altitude. The other pixels are assigned to the altitude level between these two extremes. Watershed transform is eventually applied to split the image into a set of catchment basins, where the catchment basins are formed by the regional minimum. Every pixel in the image would fall into only one of these catchment basins according to the steepest descending path. The final outcome of the watershed transform is the labeling number for each pixel of the image to a specific catchment basin [10].

Watershed transform is used for the lips segmentation as it possesses the characteristic of closed boundary segmentation. Nevertheless, the watershed transform might cause the over-segmentation problem. The total amount of catchment basins might rise up to thousands though only some of them are required. One of the over-segmentation scenarios is shown in Fig. 2.

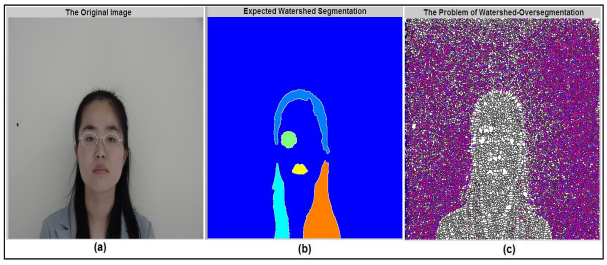

Fig. 2. (a) Original image (b) expected watershed segmentation (c) over-segmentation

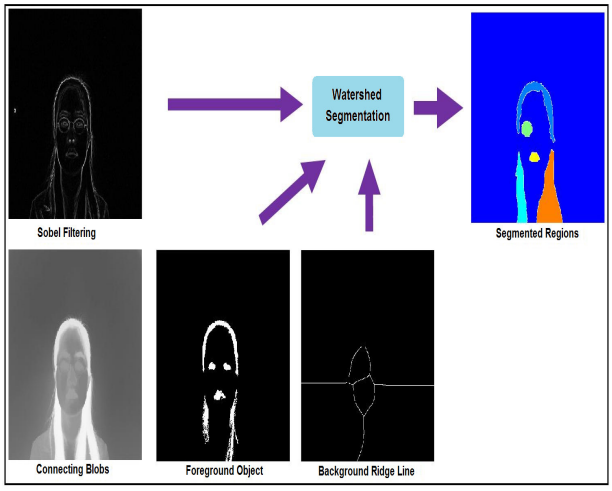

Fig. 3. The outcomes of the pre-processing procedures

As to deal with watershed over-segmentation, some pre-processing steps need to be considered instead of directly going through watershed transform. From Fig. 3, the pre-processing includes edge detection using Sobel filtering, obtaining foreground object and background ridge lines.

The segmented regions are passed to the lips detection system as depicted in Fig. 4. The cubic spline interpolant lips colour modelling which would be discussed in more details in the coming subsection is acquired to detect the lipss region from the segmented regions. If the detected region is more than one region, further symmetry detection is triggered to gain the final lips region.

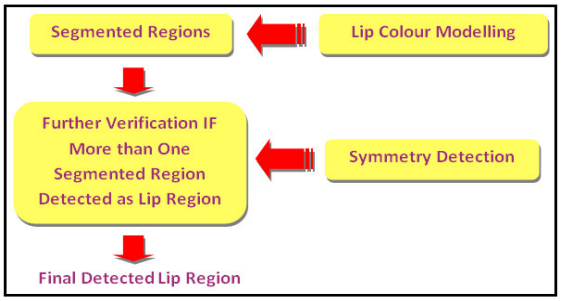

Fig. 4. The process of lips detection system

A. Lips Detection – Lips Colour Modelling

When generating the lips colour model, the images are transformed into YCbCr colour space. YCbCr colour space splits the RGB into luminance component, Y and chrominance components, Cb and Cr. As to avoid luminance issue in the system, only chrominance components, Cb and Cr are involved in lips colour feature clustering process.

6 sets of 6x6 dimensions lips area are cropped from each person in the Asian Face Database [11] and the total number of 642 sets of lips data is then plotted on a Cb-Cr graph. The plotted lips's Cb-Cr information is shown as in Fig. 5(a). Only the heavily hit pixels are considered as part of the lips colour cluster. The final lipss colour cluster after morphological closing is illustrated in Fig. 5(b).

Furthermore, the cluster boundary is created using cubic spline interpolation equated as (1)-(2) to properly encircle the cluster instead of using an incompatible triangular boundary. The cubic spline interpolant lips boundary is saved for lips detection from the segmented regions. The segmented region which belongs to the cluster is detected as the lips region.

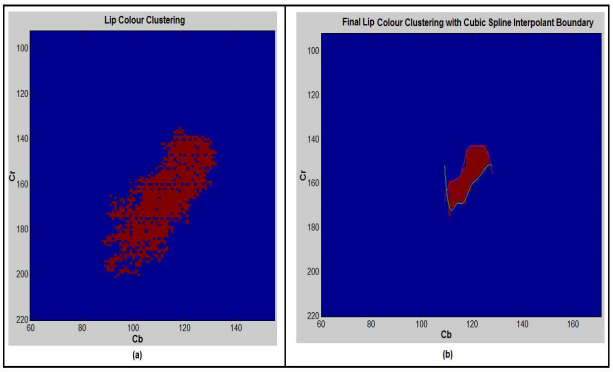

Fig. 5. Lips colour modelling (a) initial clustering (b) final clustering with cubic spline interpolant boundary

$$S(y)=\begin{cases} S_1(y) & if \quad x_1 \le x \le x_2 \\ S_2(y) & if \quad x_2 \le x \le x_3 \\ \vdots & \\ S_{k-1}(y) & if \quad x_{k-1} \le x \le x_k \end{cases} \tag{1}$$

Where S_n is the third degree polynomial which could be described as:

$$S_n(y)=a_n(y-y_n)^3+b_n(y-y_n)^2+c_n(y-y_n)+d_n \tag{2}$$

For n=1, 2, 3…k-1

B. Lips Verification - Symmetry Detection

The symmetry detection process would only be triggered, if the number of detected lips region after lips colour modelling is more than one. Only the segmented region which falls on the symmetrical axis would be considered as the lips region. The symmetry detection is performed horizontally and the equation [12] is as below:

$$HS(j)=\sum_{j=X_l}^{X_l+W}\sum_{i=Y_h}^{Y_h+HW/2}\sum_{\Delta x=1}\left|G(i,j+\Delta x)-G(i,j-\Delta x)\right| \tag{3}$$

$$j_{sym}=\arg\min_{j} HS(j) \tag{4}$$

Where $HS(j)$ is the horizontal symmetry measurement with the symmetry axis located at $x=j$.

Fig. 6 shows an example of horizontal symmetry measurement working on the input image as in Fig. 7(a). The minimum value obtained between the two peaks would be the horizontal symmetry point. After going through lips colour boundary as mentioned previously, the regions detected as lips are labelled as 3, 6 and 9 as shown in Fig. 7(b). Referring to Fig. 6, the horizontal symmetrical axis is located at x =263. Therefore, among three regions, the region which fell on the symmetrical axis, x =263, is considered as the final detected lips region.

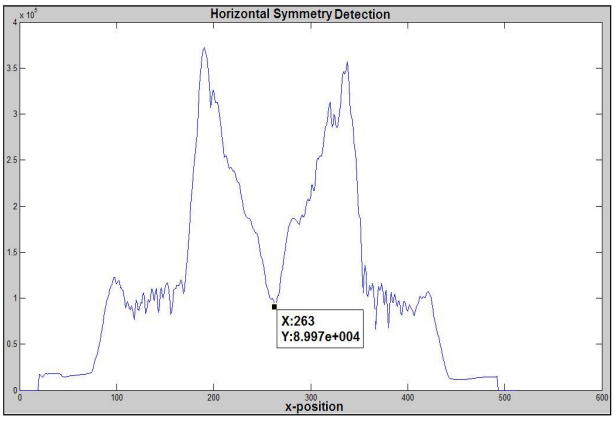

Fig. 6. Horizontal symmetry measurement

The successfully detected lips region would be forwarded to the H_∞ approach tracking system to estimate the lips location of the succeeding frames. The purpose of applying the lips tracking for the subsequent video frames instead of directly detecting the lips region for the entire video flow is to trim down the computational time of the overall image processing. After getting the predicted lips location of the next incoming frame, the watershed segmentation and detection process would only be focused on the specific area nearby the estimated point. The processing area is reduced from the entire image size to a relatively small window area. Consequently, the system efficiency would be increased. The details of H_∞ approach tracking system is discussed in the next section.

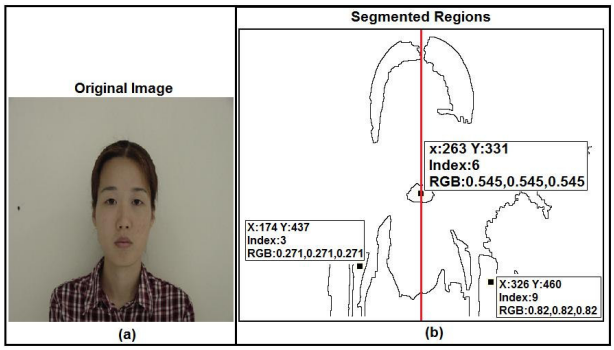

Fig.7. (a) Original image (b) lips region on the symmetrical axis

III. H_∞ Approach Lips Tracking System

A linear, discrete-time system could be mathematically expressed as:

$$\text{State equation:} \quad x_{n+1}=Fx_n+Hu_n+w_n \tag{5}$$

where x is the state of the system while F is the transition matrix that brings the state value x_n from time n to n+1; H is the matrix used to connect the input vector u with the state's variables and w is the process noise.

$$\text{Measurement equation:} \quad y_n=Rx_n+v_n \tag{6}$$

where y is the measured output, R is the observation model utilised to map the true state space to the observed space and v is the measurement noise.

The state vector comprised in the state equation (5) is the representation of the system information at a particular time step. The state vector consists of the centre position of the lips region horizontally and vertically. The state equation applied to describe the proposed lips tracking system could be mathematically written as:

$$x_{n+1}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}px_n\\ py_n\end{bmatrix}+\begin{bmatrix}t^2/2\\ t^2/2\end{bmatrix}\begin{bmatrix}ux_n\\ uy_n\end{bmatrix}+w_n \tag{7}$$

$$y_n=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}x_n+v_n \tag{8}$$

where px_n and py_n correspond to the lips position horizontally and vertically; while ux_n and uy_n are the horizontal and vertical acceleration.

A precise prediction tool is needed to successfully track the incoming lips position. Dealing with this concern, H_∞ filter that work under the worst case consideration of the estimation error is proposed as a tracking tool in this system. This filter is designed for robustness; it is applied to diminish system

modelling errors and noise uncertainty [13]. The H_∞ algorithm applied for the lipss tracking is illustrated in Fig. 7. The robustness of the H_∞ filter would only be preserved if γ is selected in the way where all the eigenvalues of P is less than one.

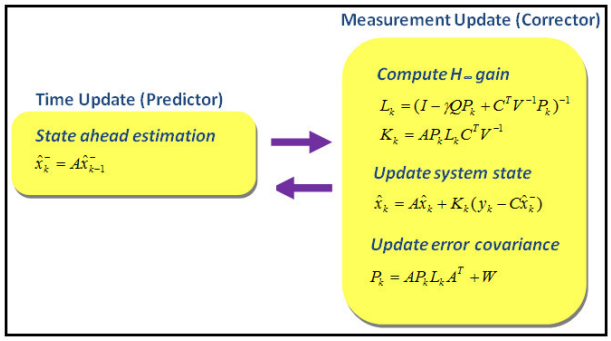

Fig. 7. The process flow of the H_∞ filter algorithm [14]

Q, W, V are the weighting matrices for the estimation error, process noise and measurement noise respectively.

IV. Simulations And Analysis

A. Simulation of Lips Segmentation using Watershed

The Asian Face Database is used to verify the proposed lips detection and segmentation system. The input image is first converted into gray level image and sent to the watershed segmentation. As mentioned in Section II, an edge image processed by Sobel filtering is needed to calculate the gradient magnitude for the watershed transform. Fig. 8(b) shows the detected edge for the incoming image. On the other hand, the input image is also applied to obtain the foreground object and the background ridge line which are depicted in Fig. 8(c) and Fig. 8(d) respectively. The watershed segmentation results are illustrated in Fig. 8(d). Compared against watershed without pre-processing, the number of segmented regions from the proposed system is tremendously reduced.

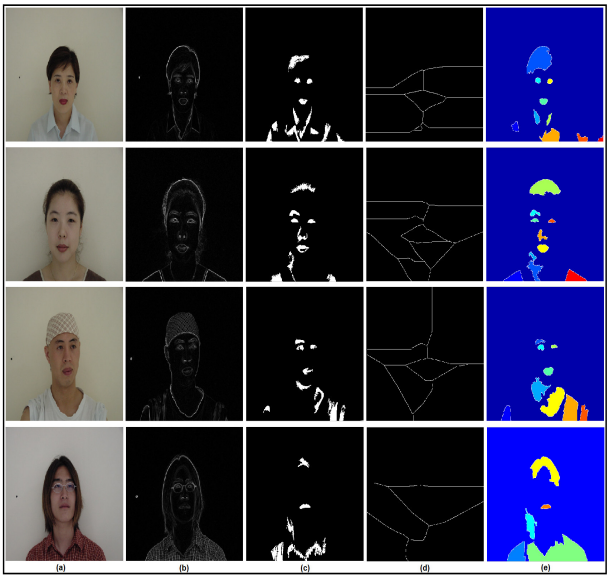

Fig. 8. (a) Input image (b) edge detection using Sobel filter (c) foreground object (d) background ridge line (e) watershed segmentation.

B. Simulation of Lips Detection using Lips Colour Modelling and Symmetry Detection

After getting the segmented regions from the watershed transform, lips detection is working on those regions to obtain the lips area. At first, the lips colour modelling is applied on the segmented region and the resultant output is shown in Fig. 9(b). If only one of the regions is detected as the lips as depicted in Fig.9(c), this particular region would be then considered as the final lips region shown in Fig. 9(d).

Fig. 9. (a) Input image (b) lips detection using lips colour modelling (c) lips region detection (d) final detected lips

Alternatively, if the detected region is more than one as shown in Fig. 10(b) and 11(b), a further verification using symmetry detection would be activated. Fig. 10(c) and 11(c) demonstrates the horizontal symmetry measurement of the input image and the minimum point between two peaks is known as the symmetrical axis. The output region from the lips colour boundary which falls on the symmetrical axis would only be verified as the final lips region as illustrated in Fig. 10(d) and Fig. 11(d). The final lips region is detected as in Fig. 10(e) and 11(e).

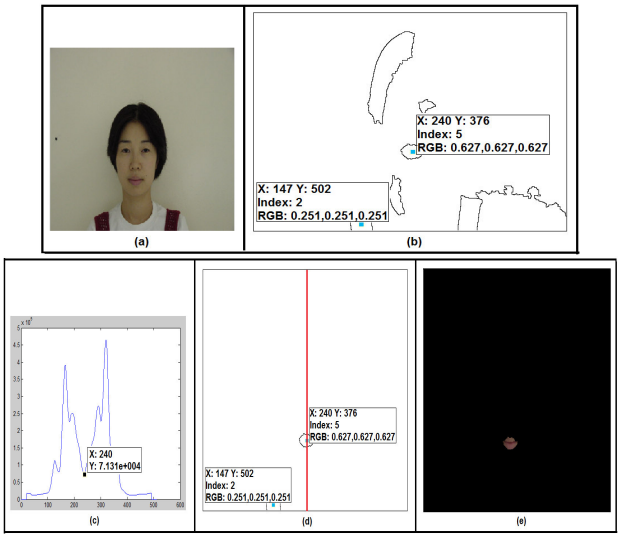

Fig. 10. (a) Input image (b) lips detection using lips colour modelling (c) horizontal symmetry measurement (d) symmetry detection (e) final detected lips

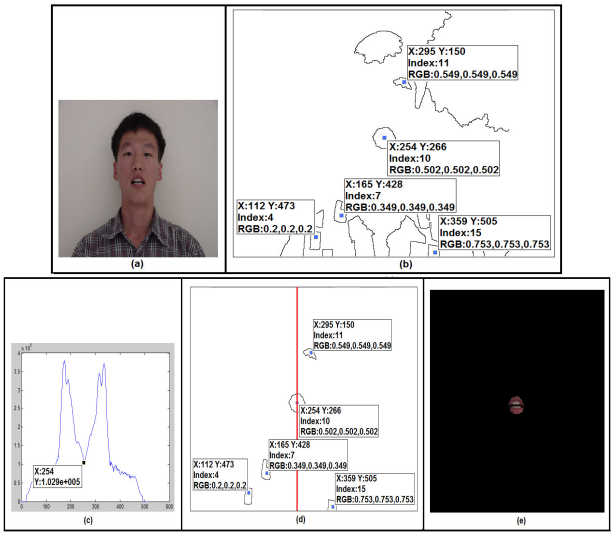

Fig. 11. (a) Input image (b) lips detection using lips colour modelling (c) horizontal symmetry measurement (d) symmetry detection (e) final detected lips

C. *Simulation of Lips Segmentation and Detection under Complex Background*

Instead of only using Asian Face Database where the background is in plain colour, the images are also obtained from [15] and in-house prepared which have the complex background are tested as well. Some of the lips detection results are shown as in Fig. 12 and Fig. 13 below:

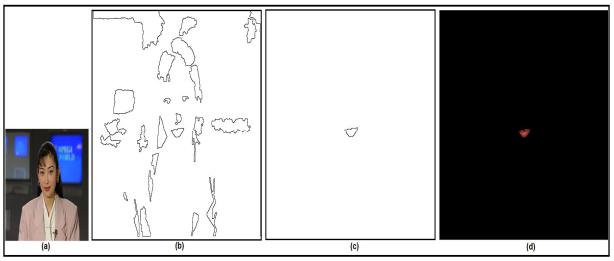

Fig. 12. (a) Akiyo image from [LS10] (b) watershed segmentation (c) lips detection using lips colour modelling (d) final detected lips

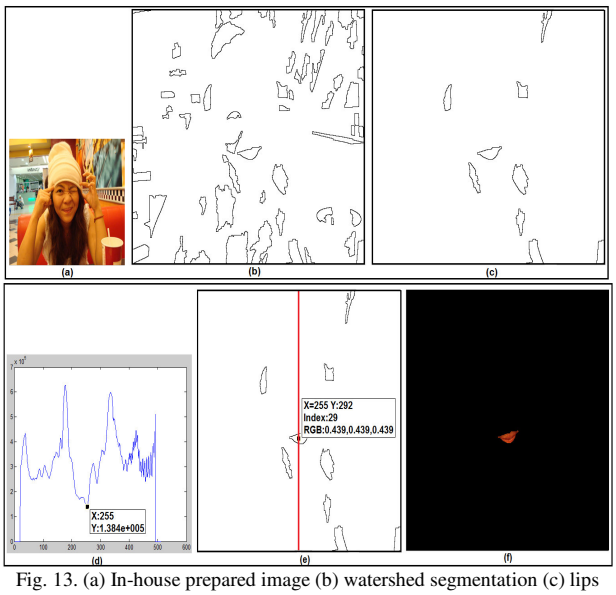

Fig. 13. (a) In-house prepared image (b) watershed segmentation (c) lips detection using lips colour modelling (d) horizontal symmetry measurement (d) symmetry detection (e) final detected lips.

D. *Evaluation of the Proposed Lips Segmentation and Detection System*

The performance of the proposed direct lips segmentation and detection system is quantitatively evaluated. The evaluation is based on percentage of overlap (*POL*) between the segmented lips region X_1 and the ground truth X_2 [15]. The ground truth is built by manually segmenting the lips region. The *POL* is equated as [15]:

$$POL = \frac{2(X_1 \cap X_2)}{X_1 + X_2} \times 100\% \tag{9}$$

According to the above measurement, the total agreement of the overlap would be 100%. Furthermore, the segmentation error (*SE*) is also evaluated using the equation as [15]:

$$SE = \frac{OLE + ILE}{2 \times TL} \times 100\% \tag{10}$$

where *OLE* is defined as the outer lips error, meaning the number of non-lips pixels being categorised as lips pixels. The *ILE*, inner lips error is explained as the number of lips-pixels being categorised as non-lips pixels. Besides, *TL* is denoted as the total lips pixels in the ground truth. The total agreement of the segmentation error would be 0%.

The *POL* and *SE* are computed for the image used in Fig. 8, Fig.12 and Fig.13. The images in Fig.8 would be named as (i)-(iv) from most top to bottom. The evaluation results are shown in Table 1.

Table 1: System evaluation based on *POL* and *SE*

Images	*POL* (%)	*SE* (%)
Fig.8 (i)	96.35	3.6074
Fig.8 (ii)	97.35	2.6445
Fig.8 (iii)	97.12	2.9016
Fig.8 (iv)	91.61	8.8776
Fig.12	90.51	9.9576
Fig.13	93.81	6.0305

From the results tabulated above, the proposed lips segmentation and detection system has an appreciable performance not only for plain background images but also for the images under uncertain circumstances.

E. *Simulation of Lips Tracking using H_∞ Approach*

After obtaining the lips location, H_∞ tracking system starts to predict the lips appearance of the incoming frame. Some of the lips tracking results are shown in Fig. 14 below.

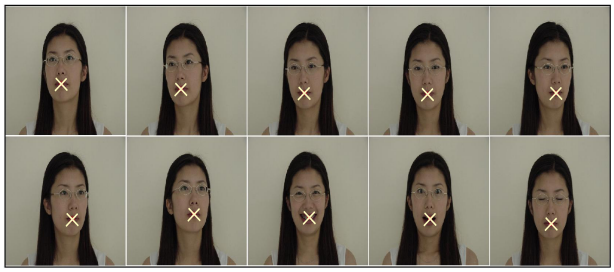

Fig. 14. Results of lips tracking

From the predicted location, the succeeding watershed segmentation and detection would only be focused on the window set around the predicted point as in Fig. 15. The computational time using H_∞ tracking approach would be reduced compared to the conventional process of segmenting and detecting the lips region from a full size image.

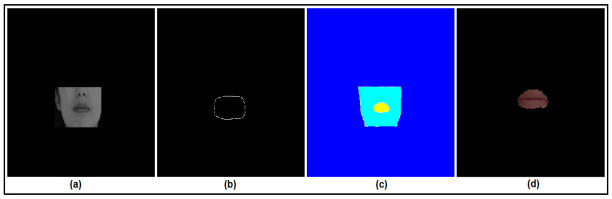

Fig. 15. The watershed segmentation and detection process in the set window around the predicted lips location

V. Conclusion

In this paper, a lips detection and tracking system based on watershed segmentation and H_∞ approach is presented. Compared to the conventional lips detection method that needs the skin/non-skin detection and face localization, the proposed system provides a potentially time saving direct lips detection technique, rendering the preliminary criterion obsolete. For the lips segmentation process, watershed transform which offers a closed boundary segmentation is applied. The final closed boundary lips region would be detected and passed to the H_∞ tracking system to track the incoming lips position. The image processing time is hence reduced as only a small window size image around the predicted location is processed to obtain the lips region instead of the entire image. From the simulation results shown, the proposed lips segmentation and detection system provides an appreciable performance not only for the plain background images but also for the images under complex background. The proposed lips detection and tracking system would be further applied into audio-visual speech recognition system in the future.

References

[1] Ara V.Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy, *Dynamic Bayesian networks for audio-visual speech recognition,* Eurasip Journal on Applied Signal Processing 2002, Vol. 2002, Issue 1, pp1274-1288

[2] Trent W.Lewis, and David M.W.Powers, *Audio-visual speech recognition using red exclusion and neural networks*, Journal of Research and Prac. In Info. Tech., Vol.35, No.1, 2003, pp41-63.

[3] Robert Kaucic, Barney Dalton, and Andrew Blake, *Real-time lip tracking for audio-visual speech recognition applications*, Proc. Of the 4th Euro. Conf. on Comp. Vis., Vol 2, pp376-387, Springer-Verlag, 1996.

[4] XiaoZheng Zhang, Charles C. Broun, Russell M. Mersereau, and Mark A. Clements, *Automatic speechreading with applications to human-computer interfaces*, Eurasip Journal on Applied Signal Processing, Vol. 2002, Issue 11, pp 1228-1247.

[5] Jian-Ming Zhang, Liang-Min Wang, De-Jiao Niu, and Yong-Zhao Zhan, *Research and implementation of a real time approach to lip detection in video sequence*, Int. Conf. on Machine Learning and Cybernetics, 2003, IEEE.

[6] Jamal Ahmad Dargham, and Ali Chekima, *Lips detection in the normalized RGB color scheme*, 2nd ICTTA 2006, Inf. And Comm. Tech., IEEE 2006.

[7] T. Wark, and S. Sridharan, *A syntactic approach to automatic lip feature extraction for speaker identification*, Proc. of the 1998 IEEE Int conf. on Acoustic, Speech, and Signal Processing.

[8] Lee Seng Yeong, Li-Minn Ang, and Kah Phooi Seng, Closed boundary face detection in grayscale images using watershed segmentation and DSFPN, Int. Sym. On Intelligent Signal Processing and Comm. Sys., 2008, IEEE.

[9] M.C. de Andrage, G. Bertrand, and A.A. Araújo, *Segmentation of microscopic images by flooding simulation: a catchment basins merging algorithm*, Conf. on Nonlinear Img. Processing, SPIE Proc., 1997, Vol.3026, pp.164-175.

[10] V.Osma-Ruiz, J.I.Godino-Llorente, N.Sáenz-Lechón, and P.Gómez-Vilda, *An improved watershed algorithm based on efficient computation of shortest path*, Pattern Recogn., Vol.40, no.3, pp.1078-1090, 2007.

[11] Intelligent Multimedia Lab, Asian Face Image Database PF01, Available: http://nova.postech.ac.kr.

[12] C. Hao-Yuan, F. Chih-Ming, and H. Chung-Lin, *Real-time vision-based preceding vehicle tracking and recognition*, in Intelligent Vehicles Symp., 2005. Proc. IEEE, 2005, pp. 514-519.

[13] Dan Simon, *From here to infinity*, Embedded Systems Programming, July 2000.

[14] Siew Wen Chin, Li-Minn Ang, and Kah Phooi Seng, *Face tracking system using H_∞ approach*, Int. Sym. On Intelligent Signal Processing and Comm. Sys., 2008, IEEE.

[15] Xiph.org video sequence test media, [Online]. Available: http://media.xiph.org/video/derf.