

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Automatic lip localization under face illumination with shadow consideration

Meng Li, Yiu-ming Cheung*

Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

ARTICLE INFO

Article history:

Received 12 November 2008

Received in revised form

11 May 2009

Accepted 21 May 2009

Available online 6 June 2009

Keywords:

Lip-reading

Lip localization

Face illumination

Shadow

ABSTRACT

Lip-reading has potential attractive applications in information security, speech recognition, secret communication and so forth. To build an automatic lip-reading system, one key issue is how to locate the lip region, particularly under the changing illumination condition. Empirical studies have shown that the recognition rate of a lip-reading system greatly relies on the accuracy of the lip localization. Unfortunately, to the best of our knowledge, lip localization under face illumination with shadow consideration has not been well solved yet. Moreover, this problem is also one of the major obstacles to keeping an automatic lip-reading system from the practical applications. This paper therefore concentrates on this problem and proposes a new approach to obtain the minimum enclosing rectangle surround of a mouth automatically based upon the transformed gray-level image. In this approach, a pre-processing is firstly made to reduce the interference caused by shadow and enhance the boundary region of lip, through which the left and right mouth corners are estimated. Then, by building a binary sequence based on the gray-level values along with the vertical midline of mouth, the top and bottom crucial points can be estimated. Experiments show the promising result of the proposed approach in comparison with the existing methods.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Motivated by human ability to lip-read, the useful information on speech content can be obtained through analyzing the subtle cue conveyed by lip movement of speakers [1]. The intimate relation between the audio and the visual sensory modality in human recognition can be demonstrated with audio-visual illusions such as the “McGurk effect” [2]. It suggests that speech perception is multimodal involving information from more than one sensory modality. In 1984, the first automatic lip-reading system was presented by Petajan [3,4]. From then on, lip-reading has received considerable attention from the community because of its potential attractive applications

in information security, speech recognition, secret communication, and so forth [5,6]. For example, we can utilize the lip-reading technique as a visual password complementary to the popular character-based password to enhance the security level in banks, e-business, home security, and so forth.

In lip-reading, one key issue is the lip localization, i.e. how to obtain the accurate position of lip or mouth from image. Paper [7] demonstrates that the error rate of this automatic visual speech recognition (AVSR) system in studio environment, i.e. ideal light condition without shadow, is 37.3%. In contrast, the visual-only word error rate will reach 76.2% when an AVSR system is utilized in real world. One main reason is that the shadows will be generated when the face illumination comes from the different directions, i.e. parts of mouth region of a speaker will be covered by shadows. Under this situation, it is non-trivial to localize the lip precisely from a face image.

* Corresponding author.

E-mail address: ymc@comp.hkbu.edu.hk (Y.-m. Cheung).



Fig. 1. The vertices marked by dot along the vertical and horizontal axis of a mouth.



Fig. 4. The filtered image composed of two filtered sub-images.

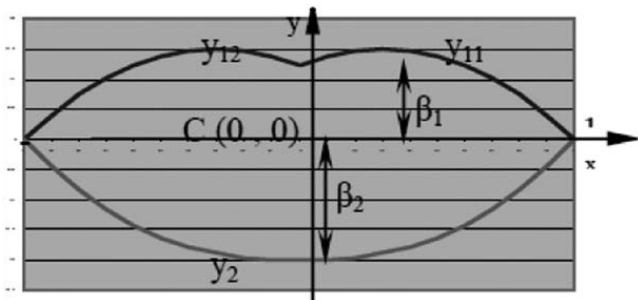


Fig. 2. The illustration of lip shape model proposed in [11].

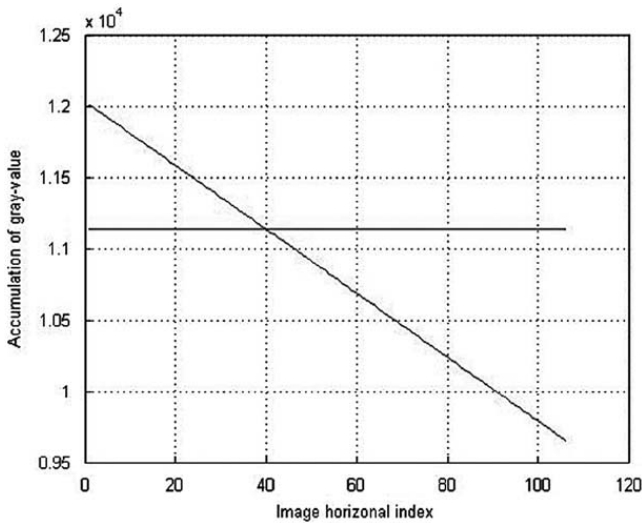


Fig. 3. The linear fitting of accumulation of gray-level value for each column and the mean value. The intersection of the two curves is corresponding into the boundary of shadow.

Consequently, the imprecise localization of lip will mostly lead to the degraded performance of an AVSR system. This implies that the accuracy of lip localization is one of the most important factors to determine the recognition rate of lip-reading. In this paper, we will therefore concentrate on studying the lip localization only.

Thus far, several methods have been proposed to enhance the performance of lip localization for AVSR system. For instance, paper [8] presents an approach that employs the hue-filter to distinguish lip and surrounding



Fig. 5. The image using the pre-processing that is made up by the two components: I_{sl} and I_{sr} . The vertical crease is corresponding into the shadow boundary.

skin region. The papers in [9–12] utilize the information of red component and saturation to localize the lip region. Also, paper [13] utilizes a gradient-based Canny edge detector to locate the mouth corner. In [14], the input image is projected into YUV color coordinate system and the accumulations of V value in each row and column of the image are utilized to estimate the crucial points (i.e. the top, bottom and two mouth corners) of a lip.

Furthermore, a class of widely used methods is active shape model (ASM) [15] or active appearance model (AAM) [16]-based ones [17–23]. They build a deformable model for lip by learning the patterns of variability from a training set of correctly annotated images. The shape of modal can be adjusted by a parameter set so as to match and locate the lip in test image. Empirical studies have shown their success, but they need to label some landmarks manually for training. Alternatively, optical flow-based methods give an effective way to locate the lip region [24]. They utilize the apparent velocity distribution of the brightness patterns in an image to obtain the boundary of lip. The optical flow-based methods can give the important information regarding the spatial arrangement of the viewed objects and the rate of change in this arrangement, but sensitive to the translation, scaling, rotation, and the change of illumination condition [25] in particular.

In general, the methods stated above make the lip localization in a studio environment only, but not applicable to a shadow situation, in which the boundary between lip and surrounding skin region, especially the area near mouth corners, cannot be distinguished precisely. In fact, under the circumstances, these methods may not locate the region of interesting (ROI) accurately,

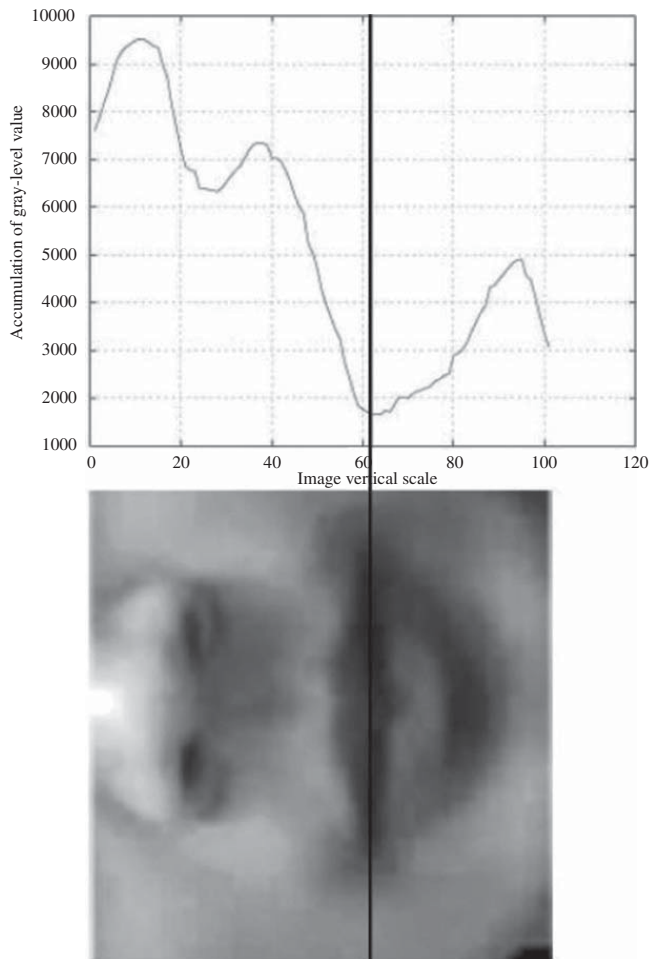


Fig. 6. Accumulation curve of gray-level value for each row. The vertical line crossing the curve and lip image represents the relation between the horizontal midline of mouth and the minimum value of the accumulation curve.

thus degrading the subsequent recognition accuracy. In the literature, some shadow detection methods have been proposed, e.g. see [26–28], which can detect a shadow region and eliminate it then. The underlying techniques are applicable for video segmentation in indoor environment, but they have not been studied for lip localization yet. Alternatively, to circumvent the shadow effect, some approaches need to make the landmarks around mouth, e.g. see [29,30]. Nevertheless, to the best of our knowledge, the lip localization under the shadow environment has not been well solved yet in the literature.

In this paper, we focus on lip localization under the face illumination with shadow consideration. The geometric lip features of interest are four outer lip crucial points, as marked by dot in Fig. 1, along the horizontal and vertical directions. We propose an approach to extract the minimum enclosing rectangle of lip automatically based upon the gray-level image. This approach utilizes the mean filter and the image transformations to circumvent the noise caused by shadow. Subsequently, the crucial points of the ROI are estimated via analyzing the curve of gray-level values along with the vertical and horizontal midline of mouth, respectively. Experiments have shown the promising result of the proposed approach in comparison with the existing methods.

The remainder of this paper is organized as follows. Section 2 overviews the two existing typical lip localization methods, which will be compared with our proposed method in Section 4. Section 3 presents our new approach for lip localization under the shadow environment. In Section 4, we will conduct the experiment to empirically compare the proposed approach with the existing methods. Finally, we draw a conclusion in Section 5.

2. Overview of existing lip localization methods

Various works have been made to locate lip position for the application of AVSR in the last decade. In this section, two typical approaches will be reviewed.

2.1. Method 1

Salah Werda et al. [11] proposed an algorithm for automatic lip and point of interest localization on a

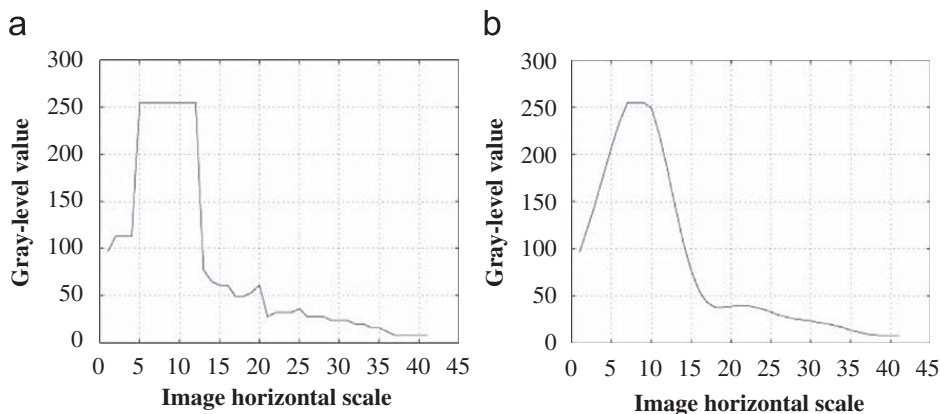


Fig. 7. (a) The curve of original gray-level value, represented by G , along with the horizontal midline of mouth, and (b) the corresponding filtered gray-level value, represented by G_f .

speaker's face based upon the color information of mouth. The main procedures will be summarized in the following. Interested readers may refer to [11] for more details.

Firstly, given the RGB values of a mouth image, this approach reduces the lighting effect by utilizing the color system conversion:

$$R_n = 255 \frac{R}{Y}, \quad G_n = 255 \frac{G}{Y}, \quad B_n = 255 \frac{B}{Y}, \quad (1)$$

where Y is the intensity component value calculated by

$$Y = 0.299R + 0.587G + 0.114B. \quad (2)$$

Then, a binary threshold based on the R_n value is utilized, knowing that R_n is the most dominant component in lip region.

After the binarization step, an oily filter is utilized in the image. This filter replaces the pixel at (x,y) with the value that occurs most frequently in the region of filter window. In the filtered image, the black region is considered as the ROI.

For the ROI in source image, saturation value of each pixel is calculated, and accumulation for each row can be

obtained then. The corresponding row is regarded as the darkest axis. Along with this axis, a scanning is utilized to localize the local maxima saturation values via the following equation:

$$S_c(x,y) = \alpha \nabla I(x,y) + \beta S(x,y), \quad (3)$$

where $\nabla I(x,y)$ is the gradient pixel value, and $S(x,y)$ is the saturation component value. Furthermore, α and β are fixed with $\alpha + \beta = 1$. Extrema of these detected local maxima pixels will be defined as the left and the right corners of the mouth.

To extract the external lip contour, a model is built using the following equations:

$$y_{11} = -\frac{-\beta_1 + 2\alpha_1\beta_1 + x^2\beta_1 + 2x\alpha_1\beta_1}{-2\alpha_1 + \alpha_1^2 + 2x\beta_1\varepsilon_1 + 2\beta_1\alpha_1\varepsilon_1 + 1}, \quad (4)$$

$$y_{12} = \frac{\beta_1 + 2\alpha_1\beta_1 + x^2\beta_1 + 2x\alpha_1\beta_1}{-2\alpha_1 + \alpha_1^2 + 2x\beta_1\varepsilon_1 - 2\beta_1\alpha_1\varepsilon_1 + 1}, \quad (5)$$

$$y_2 = \frac{\beta_2 - \beta_2(x^2)^{\alpha_2}}{2\alpha_2\beta_2\varepsilon_2 - 1}, \quad (6)$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$, and $\varepsilon_1, \varepsilon_2 \in [-1, 1]$ are the parameters that are fixed experimentally to construct the shape of lip model. Furthermore, Eqs. (4) and (5) describe the right and left higher sub-model (i.e. y_{11} and y_{12}), respectively, and Eq. (6) describes the lower sub-model (i.e. y_2). For $x \in [-1, 1]$ with the origin at $C(0,0)$, the interpretations of y_{11}, y_{12} and y_2 are illustrated in Fig. 2.

For each image, optimum lip model is selected via the extern energy of the model. This energy is based upon the gradient of an image. The extern energy for the upper and lower lip model is shown below:

$$E_{ext}(L_{Mod}) = \sum_{S=1}^n |\nabla I(s)|, \quad (7)$$

where L_{Mod} is the lip model, and n is the number of points in the lip model. Consequently, we vote for upper and lower lip models, which have the higher $E_{ext}(L_{Mod})$.

2.2. Method 2

Another simple but effective automatic lip localization is proposed in [14]. In this approach, the RGB values of a mouth image are projected into YUV color coordinate system. The coordinate V is calculated based upon

$$V = 0.615R - 0.515G - 0.1B, \quad (8)$$

through which the coordinate V value images are utilized for further processing.

In this approach, the crucial points of lip are the same as our proposed method described later. To find the position of these points, an accumulation of the V values for each row and each column of the image is made. In this way, two accumulation curves for each image can be obtained. Then, a 5-point derivative filter shown below is employed in each curve:

$$DD(c) = A(c-2) + 2A(c-1) + 2A(c+1) + A(c+2), \quad (9)$$

where $A(c)$ is the accumulated value of row or column number c , and $DD(c)$ is the resulting derivative value.

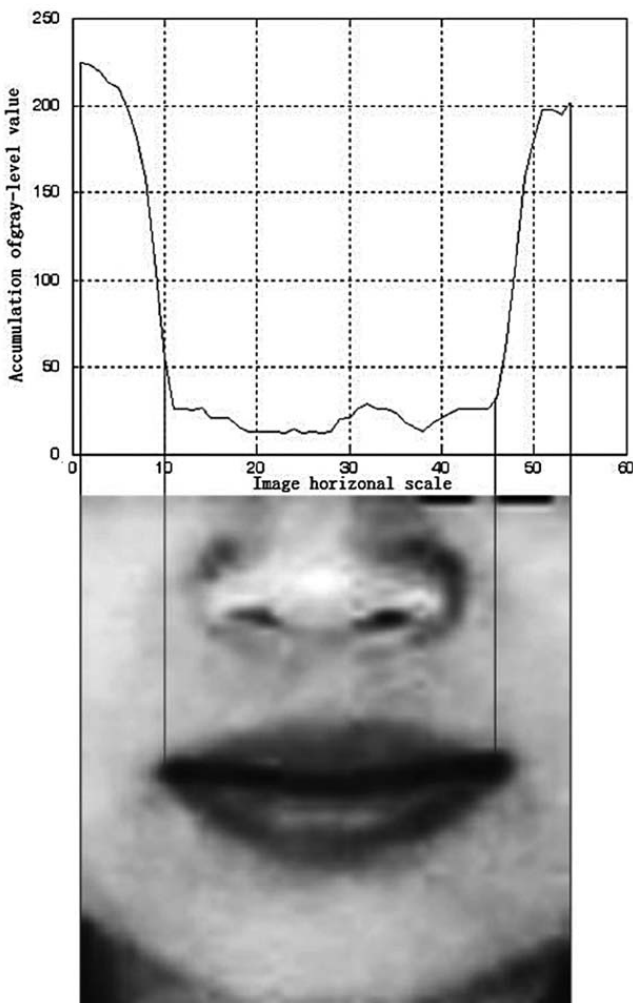


Fig. 8. Lip image under ideal illumination condition and the relevant curve of gray-level value along with the midline of the mouth, namely the whole curve consisting of G in both left and right sub-images divided by the shadow boundary. It can be seen that the positions of mouth corners correspond to the steep slopes of curve.

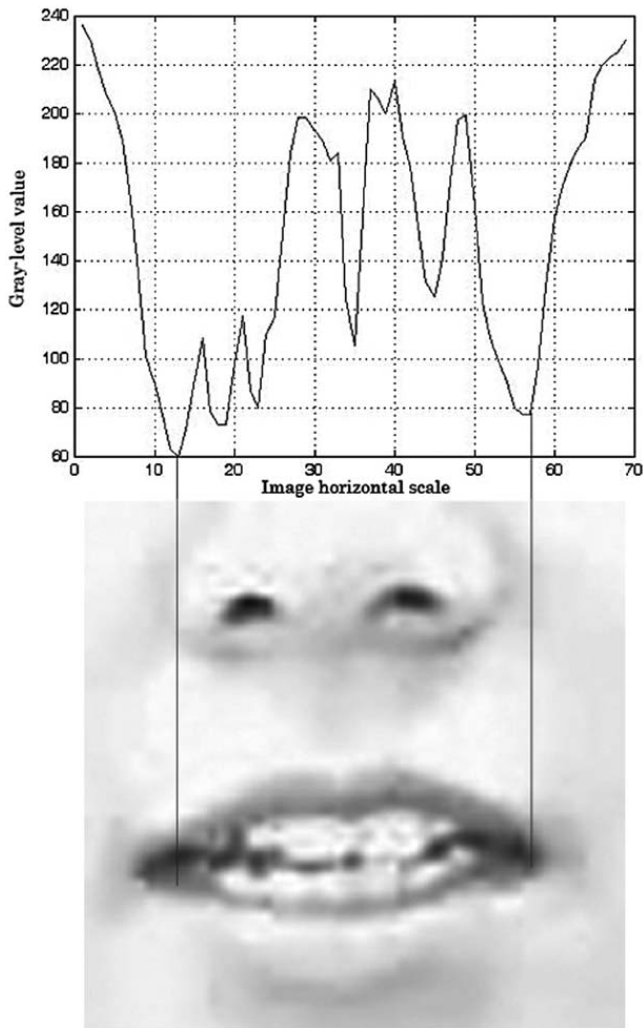


Fig. 9. Lip image with the teeth visible. It is clear that the teeth region corresponds to the segment of anomalous wave.

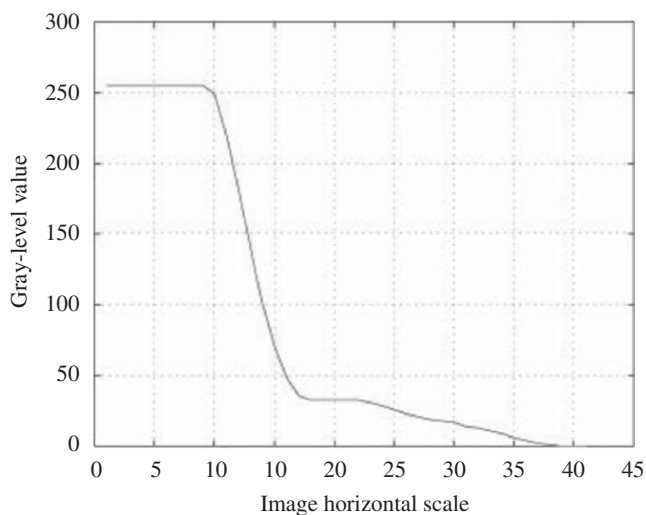


Fig. 10. The result of normalized G_m .

The maximum and the minimum points on the derivative y-axis curve are retained as the row position of the top and bottom crucial points, respectively. Similarly, the derivative curve for the x-axis accumulation

gives the column position of left and right crucial points, say the mouth corner.

3. The proposed lip localization method

The lip localization we propose is performed in two stages. The first stage considers the estimation of left and right mouth corners in each frame of a given video sequence. In the second stage, vertices in vertical direction are extracted based upon the positions of mouth corner.

Before showing the details, the definition of white point is needed. Specifically, this paper utilizes the D50 white point to denote the white color in *LAB* color space.

The images captured by camera are composed of RGB values. We heuristically project these RGB values into the gray-level space based on the following equation:

$$I = 0.299R + 0.587G + 0.114B. \quad (10)$$

From the practical viewpoint, it is inevitable that some noise arises from shadow. To reduce the interference brought by shadow, we calculate the accumulation of the gray-level value for each column of the image, and obtain column index corresponding to the mean value of the accumulation curve as the boundary of shadow as shown in Fig. 3.

Since the gray level of shadow and mouth area are represented by the close contrast values, a contrast stretching adjustment is performed in each two sub-images divided by the shadow boundary to enhance the contrast between lip and surrounding skin region according to the following equation:

$$I_e = \frac{255(I - I^{min})}{I^{max} - I^{min}}, \quad (11)$$

where I^{min} is the minimum gray-level value in the image, and I^{max} is the maximum one. Subsequently, a 3×3 mask M with

$$M = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{bmatrix}$$

is utilized to perform mean filter in these two sub-images. That is,

$$I^{(i+1)}(x, y) = \sum_{s=-1}^1 \sum_{t=-1}^1 I^{(i)}(x + s, y + t)M(s, t), \quad (12)$$

where $I^{(i)}$ is the result of the i th time filter. The times that the filter is performed are determined by

$$\delta_i = dist(I^{(i+1)}, I^{(i)}), \quad (13)$$

where δ_i is the Euclidean distance between $I^{(i)}$ and $I^{(i+1)}$. The procedure should be stopped as soon as δ_{i+1} is greater than or equal to δ_i , and then $I^{(i)}$ is marked as I_f . In our approach, I_f for each sub-image is calculated individually, and build a whole one. A snapshot of I_f is shown in Fig. 4.

Then, the subtracted image between $I^{(0)}$ and I_f can be calculated and marked as I_s . To get a clear view of its effects, an image inversion transformation is employed, and its result is shown in Fig. 5. Moreover, to reduce the

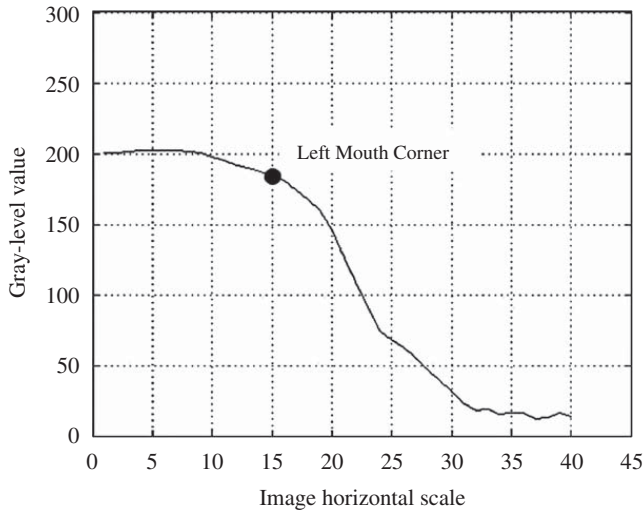


Fig. 11. An example in which the mouth corner point is in the curve segment that has the significant descending trend, but not have the maximum derivative.

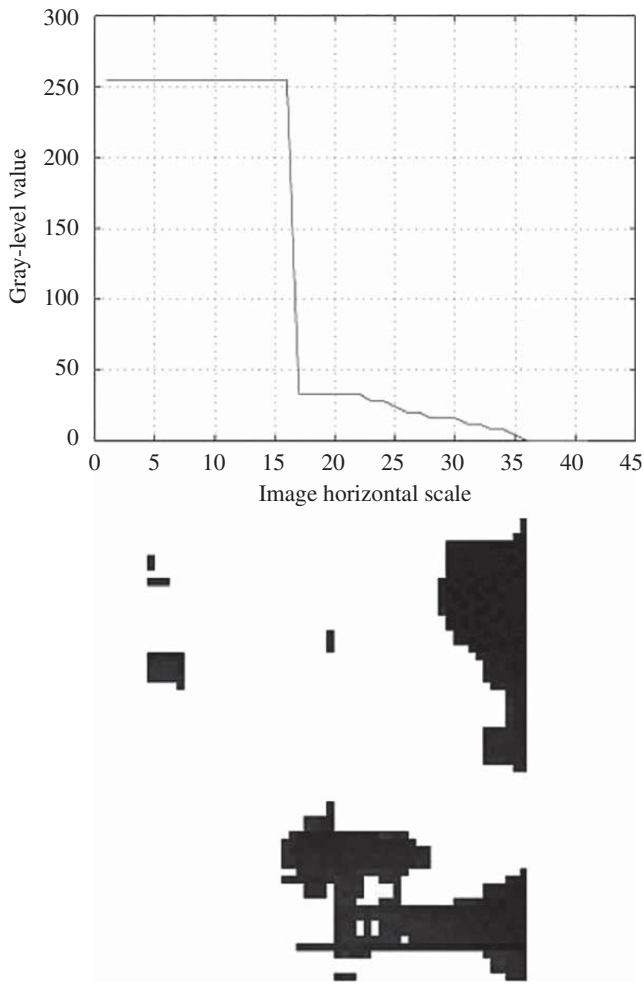


Fig. 12. Final result of gray-level curve along with the midline and the corresponding image in the extraction procedure. It can be seen that the left mouth corner corresponds to the first inflexion of the curve.

effect caused by shadow, I_s is divided into two parts by the shadow boundary that is marked as I_{sl} and I_{sr} .

3.1. Horizontal vertices localization

To start, we should find the horizontal midline of mouth. We make an accumulation of gray-level value for each row of the image as shown in Fig. 6. The slopes of the curve contain information regarding the boundaries between lips and surrounding skin region. The minimum value on the curve is retained as the row position of mouth corner points, and the row can be named as the horizontal midline of mouth. This midline can be shared by I_{sl} and I_{sr} . The following discussion in this subsection is based on I_{sl} , and the right corner can be extracted in a similar way.

The curve of gray-level values along with the horizontal midline is saved in the vector G , i.e. G is a representation of the curve. However, in this curve, there are a lot of high frequency noises caused by complexion, moustache and so on. To circumvent these high frequency noises, a low pass Butterworth filter is utilized, in which passband corner frequency is 10 Hz, stopband corner frequency is 20 Hz, passband ripple is 3, and stopband attenuation is 20. The filtered curve is marked as G_f . G and G_f are shown in Fig. 7.

Ideally, the shape of gray-level value curve along with the midline is shown in Fig. 8. The ideal shape of gray-level value curve along with the horizontal midline of I_{sl} should be monotonically decreased. The point with the maximum absolute slope may correspond to the left mouth corner because of the gray-level difference between the skin and the lip color. However, the practical curve is usually anomalous, which may be caused by the illumination condition or visibility of teeth (see Fig. 9). In our approach, the curve G_f is arranged so as to reduce the noises stated above and highlight the corner position via the following procedure.

We obtain the index of the first minimum from the left marked as m . Then, we utilize the following equations to make the curve monotonic and save it into a new n -dimensional vector named G_m , whose i th element is defined as

$$G_m^{(i)} = \begin{cases} G_f^{(i)} & (G_f^{(i)} \geq G_f^{(i+1)}), \\ G_m^{(i+1)} & (G_f^{(i)} < G_f^{(i+1)}), \end{cases} \quad (14)$$

where $i = m - 1, m - 2, \dots, 1$, and $G_f^{(i)}$ is the i th element of vector G_f . Further, we let

$$G_m^{(i)} = G_f^{min} \quad (15)$$

as $i = m, m + 1, \dots, n$. The initial value of G_m is

$$G_m^{(m)} = G_f^{(m)}. \quad (16)$$

Furthermore, G_m is projected into the range between 0 and 255 by the following equation:

$$C = \frac{255(G_m - G_m^{min})}{G_m^{max} - G_m^{min}} \quad (17)$$

as shown in Fig. 10.

Nevertheless, in some cases, the point may not have the maximum derivative (see Fig. 11) although the left

mouth corner may be on a segment curve of C which has the significant descending trend.

Thus, an iteration procedure is employed to make the segment that includes the left mouth corner steep so as to be distinguished easily. Firstly, the average can be calculated by

$$c^{avg} = \frac{\sum_{i=1}^k C^{(i)}}{k}. \quad (18)$$

The following equation is then utilized to adjust the

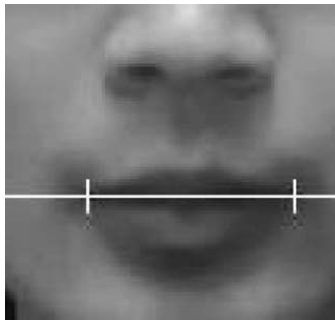


Fig. 13. The estimate of crucial points in the horizontal direction.

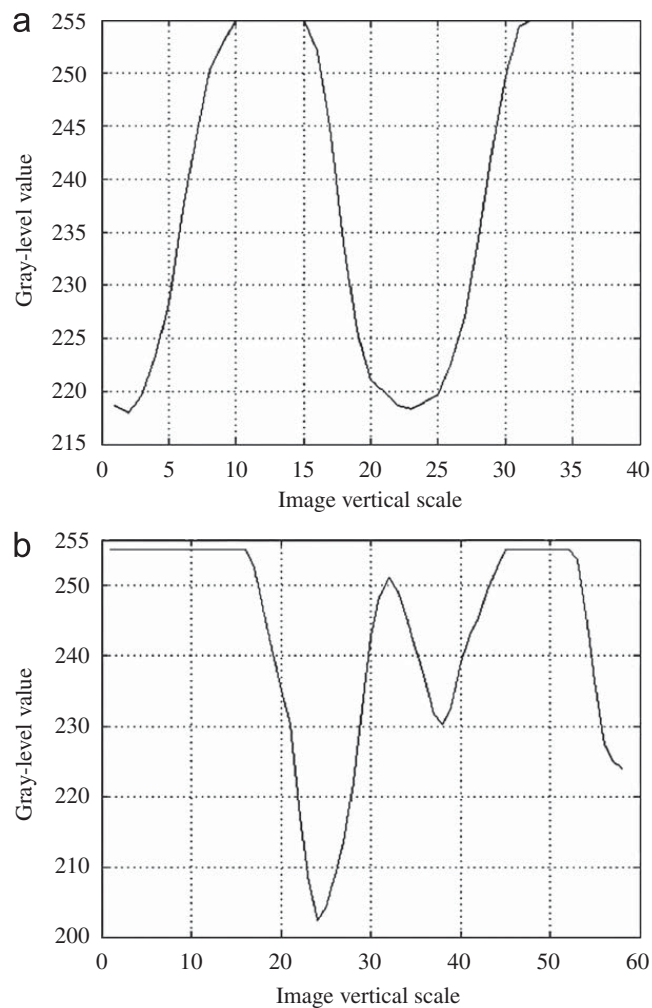


Fig. 14. The curve of (a) G_{m1} and (b) G_{m2} .

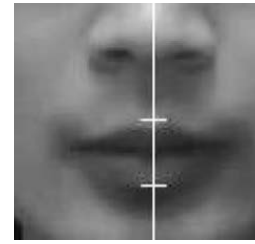


Fig. 15. The estimate of crucial points in normal direction.

contrast of image:

$$I_{out} = \begin{cases} 255 & (1.5c^{avg} < I_{in} < 1), \\ \frac{500}{c^{avg}} - 500 & (0 < I_{in} \leq 1.5c^{avg}), \end{cases} \quad (19)$$

where I_{in} is the input gray-level value, and I_{out} is the output.

For the adjusted image, an 11×1 searching block is performed along with the midline, positions of the most left and right non-pure white block are marked as the column of mouth corner candidates. Repeat the extraction steps above until the position of mouth corner candidates is unchanged any more or the image becomes a binary one. Fig. 12 illustrates the final result of image and curve C . Fig. 13 shows the estimate of the horizontal crucial points, i.e. corners of mouth.

3.2. Normal crucial point localization

Based upon the positions of left and right corners of a mouth estimated in Section 3.1, the center point of mouth can be calculated easily. The column index of the center point is marked as the vertical midline of mouth.

We obtain the gray-level values along with the vertical midline of mouth in I_s . The points with the row between the image top and the center point of mouth compose a vector named G_{m1} . The points with the row between the center point of mouth and image bottom compose a vector named G_{m2} . The curves of G_{m1} and G_{m2} are shown in Fig. 14.

For these two curves, the points performing the extreme value of maximum are marked as 1, and the points performing the extreme value of minimum are marked as 0. Then, two binary values named B_1 and B_2 are obtained. We let the two binary values B'_1 and B'_2 be

$$\begin{aligned} B'_1 &= B_1 \oplus (B_1 \ll 1) \\ B'_2 &= B_2 \oplus (B_2 \ll 1), \end{aligned} \quad (20)$$

where " \ll " is the left circle shift operator.

For B'_1 , the point corresponding to the first appearance of "1" is retained as the row of vertical vertices of the outer upper lip. For B'_2 , the point corresponding to the last appearance of "1" is retained as the row of the vertical vertices of the outer lower lip. Hence, the normal crucial points are extracted as shown in Fig. 15.

Consequently, we can obtain the minimum enclosing rectangle as shown in Fig. 16.

In summary, the procedure of the proposed approach is given as follows:

Step 1: Convert the original lip image into gray-level space.

Step 2: Project the gray value onto the x -axis, and get the mean value of the accumulation curve which is marked as y_{mean} . Use a linear function $y = kx + b$ to fit the accumulation curve, and calculate the shadow boundary $x = (y_{mean} - b)/k$.

Step 3: Split the gray-level image into the two sub-images based upon the shadow boundary, and Steps 2–10 are implemented in the two sub-images, respectively.

Step 4: Employ the contrast stretching and 3×3 mean filter by Eqs. (11) and (12).

Step 5: Repeat Step 4 until the Euclidean distance between the image before filtering and the one after filtering (see Eq. (13)) is not descending, the final image is marked as I_f .

Step 6: Get the subtracted image, I_s , between $I^{(0)}$ and I_f .

Step 7: Project the gray-level value of I_s onto y -axis, the row index of image corresponding into the minimum is selected as the horizontal midline of mouth.

Step 8: Smooth, filter, and normalize the gray-level value curve along with the horizontal midline of mouth by Eqs. (14)–(17).

Step 9: Adjust the contrast of image by Eqs. (18) and (19).

Step 10: Search along with the midline via an 11×1 block, the column corresponding into the most left non-pure white block is selected as the row of crucial point in horizontal.

Step 11: Get the center point of mouth based on the left and right mouth corner.

Step 12: The gray values between the top point of vertical midline and center point of mouth in I_s compose a vector named G_{m1} , and the gray values between the bottom point and center point compose a vector named G_{m2} .

Step 13: Based on G_{m1} and G_{m2} , we get the two binary sequences: B'_1 and B'_2 by Eq. (20). The point corresponding to the first appearance of '1' in B'_1 is the top crucial point, and the point corresponding to the last appearance of '1' in B'_2 is the bottom crucial point.

It can be seen that the time complexity of this approach is $O(n)$.

4. Experiment

4.1. Database

To demonstrate the performance of the proposed approach in comparison with the existing Methods 1 and 2 as described in Section 2, we established an image database, in which the condition of video capturing is grouped into two classes:

- (1) low contrast environment and
- (2) variational shadow.

For the first category, video clips were acquired by the system that captures a 128×128 window at 25 frames per second and 24-bit pixel resolution. The video clips were captured with all artificial lighting equipments, but only the daylight was utilized. For the second category, based on the conditions of the first category, two 36 W fluorescent lamps were complemented, where they were placed at the left and right side on top of the speaker, respectively. Furthermore, a 7 W warm white light was utilized in 10 different positions so as to make variational shadow effect as shown in Fig. 17.

To test the robustness of the approaches, all of the videos were captured by low resolution camera, e.g. camera in notebook or mobile phone. Some sample frames in the database are shown in Fig. 18.

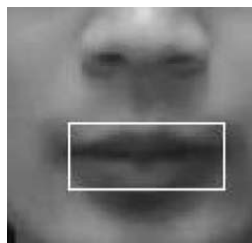


Fig. 16. The minimum enclosing rectangle of mouth.

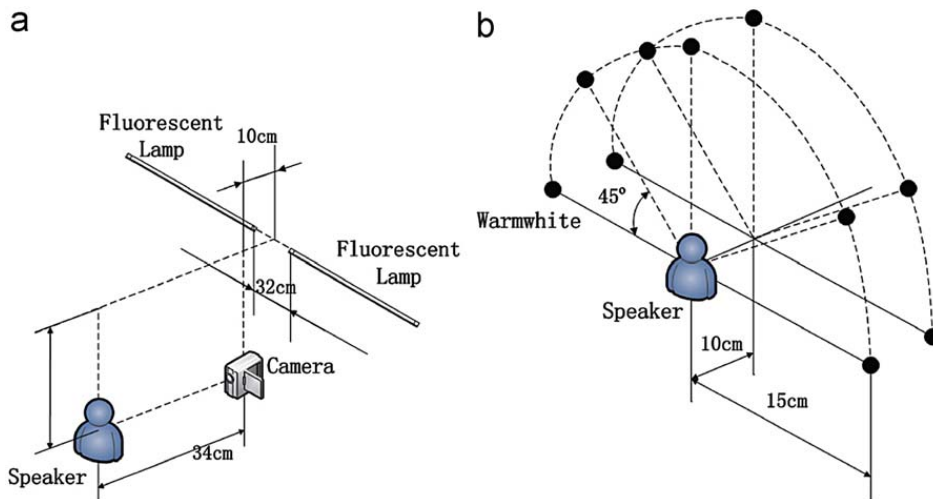


Fig. 17. (a) The environment of video clip capturing and (b) the testing conditions of illumination.

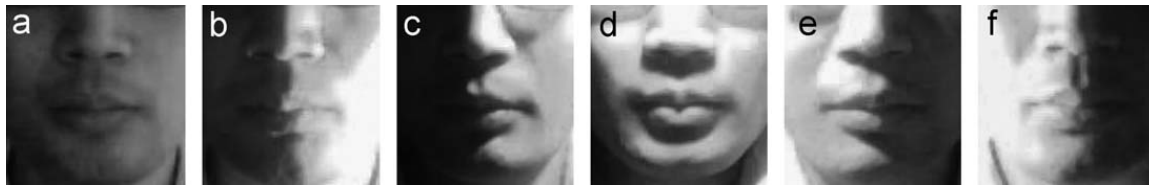


Fig. 18. Sample frames in database with (a) low contrast, (b) left side light, (c) upper left side light, (d) top light, (e) upper right side light, and (f) right side light.

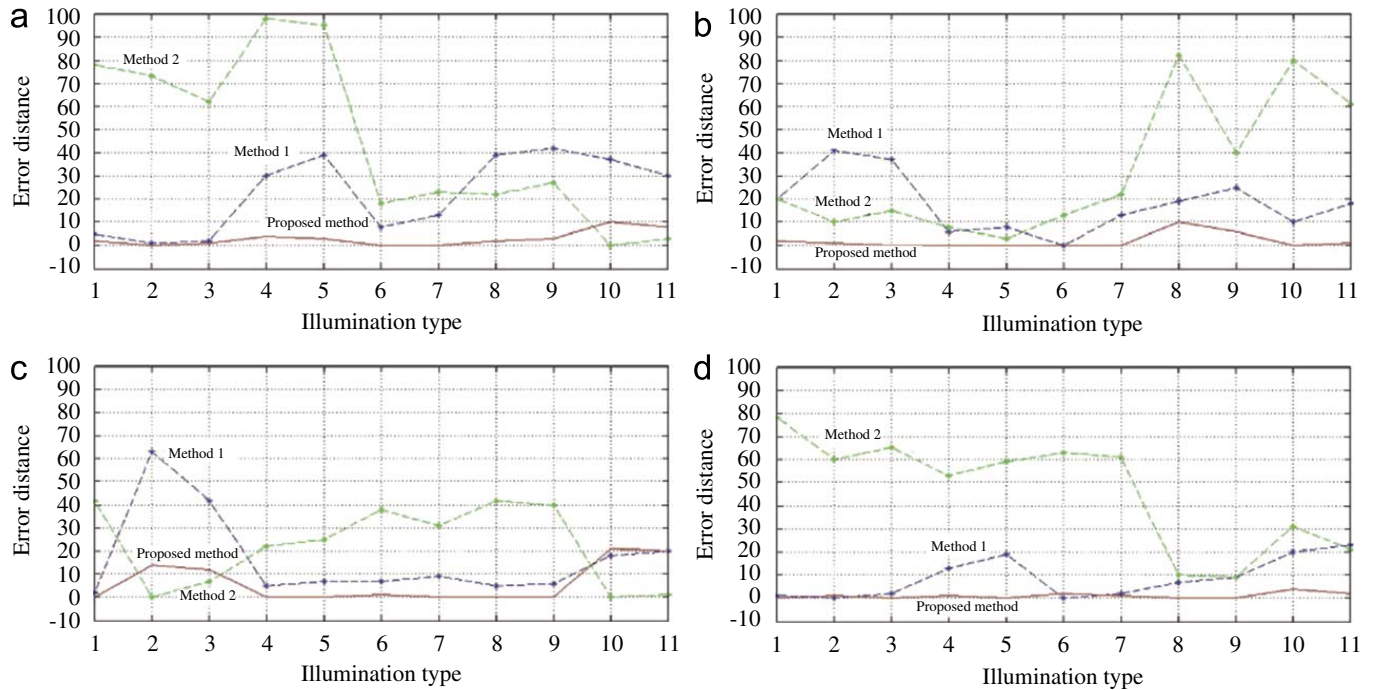


Fig. 19. The error curves of four crucial points: (a) left corner point, (b) right corner point, (c) top point, and (d) bottom point. The horizontal direction axis represents the different 11 light conditions, while the vertical one represents error pixels. The solid line is the result obtained by the proposed approach. Correspondingly, the two dashed lines are the results obtained by Method 1 [11] and Method 2 [14], respectively.

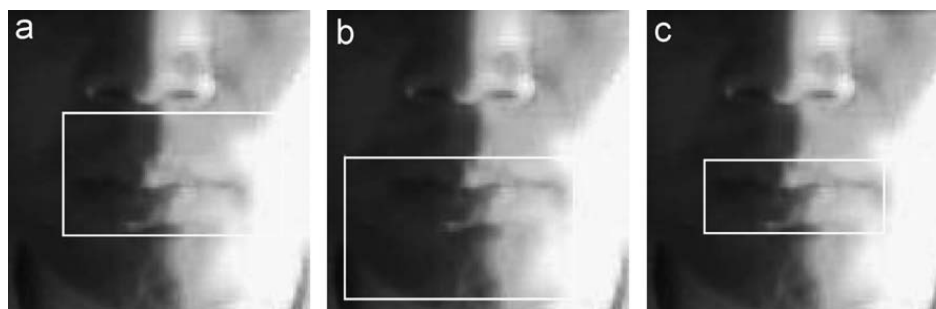


Fig. 20. The lip location results obtained by (a) Method 1 [11], (b) Method 2 [14], and (c) the proposed approach.

4.2. Experiment results

We investigated the proposed approach on the database described above. The database involved 10 persons (8 males and 2 females). For each speaker, there were 11 image groups (10 illumination type stated above and a type without warm white lamp on) classified by light condition as shown in Fig. 18. We applied the approach to each image and estimated the positions of four crucial points. To evaluate the accuracy of our

approach, we marked the crucial points in each image manually. Subsequently, the location error was calculated by

$$err = \sqrt{(x_{est} - x_{dat})^2 + (y_{est} - y_{dat})^2}, \quad (21)$$

where x_{est} and y_{est} are the estimated position of crucial point, and x_{dat} and y_{dat} are the position of corresponding mark. For the four crucial points, we can obtain four errors, respectively.

For comparison, Method 1 in [11] and Method 2 in [14] were also applied to the same images. Fig. 19 shows their error curves on four crucial points, respectively. Fig. 20 shows a snapshot of the experimental results. Specifically, the mean and standard deviation of the absolute errors are utilized to describe the accuracy and stability robustness of the three approaches quantitatively. The experiment result is shown below.

	Method 1	Method 2	Proposed approach
Mean	16.4318	36.6136	3.0000
Standard deviation	15.1049	28.5447	5.1984

It can be seen that the proposed approach outperforms the other two existing methods.

5. Conclusion

In this paper, we have proposed a new approach to automatic lip localization via obtaining the minimum enclosing rectangle surround of mouth automatically based upon the gray-level image. This approach features high accuracy of lip localization and robust performance against the shadow caused by illumination from the different directions. Experiments have shown the promising result of the proposed approach in comparison with the existing methods.

Acknowledgment

The work described in this paper was fully supported by the Faculty Research Grant of Hong Kong Baptist University with the Project code: FRG/07-08/II-88.

References

- [1] J. Bulwer, *Philocopus, or the Deaf and Dumbe Mans Friend*, Humphrey and Moseley, 1648.
- [2] H. McGurk, J. McDonald, Hearing lips and seeing voices, *Nature* 264 (1976) 746–748.
- [3] E. Petajan, Automatic lipreading to enhance speech recognition, Ph.D. Thesis, University of Illinois, 1984.
- [4] E. Petajan, Automatic lipreading to enhance speech recognition, in: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 1985, pp. 40–47.
- [5] T. Chen, R. Rao, Audio-visual integration in multimodal communication, *Proceedings of the IEEE* 86 (5) (1998) 837–851.
- [6] G. Potamianos, C. Neti, J. Luetttin, I. Matthews, Audio-visual automatic speech recognition: an overview, in: G. Bailly, E. Vatikiotis-Bateson, P. Perrier (Eds.), *Issues in Visual and Audio-Visual Speech Processing*, MIT Press, Cambridge, MA, 2004.
- [7] G. Potamianos, C. Neti, Audio-visual speech recognition in challenging environments, in: *Proceedings of European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003, pp. 1293–1296.
- [8] T. Coianis, L. Torresani, B. Capril, 2D deformable models for visual speech analysis, in: *NATO Advanced Study Institute*, Springer, Berlin, 1995, pp. 391–398.
- [9] G. Potamianos, H. Graf, E. Cosatto, An image transform approach for HMM based automatic lipreading, in: *Proceedings of IEEE International Conference on Image Processing*, Seattle, WA, 1998, pp. 173–177.
- [10] A.W. Liew, S. Leung, W. Lau, Segmentation of color lip images by spatial fuzzy clustering, *IEEE Transactions on Fuzzy Systems* 11 (4) (2003) 542–549.
- [11] S. Werda, W. Mahdi, A. Hamadou, Colour and geometric based model for lip localisation: application for lip-reading system, in: *Proceedings of IEEE International Conference on Image Analysis and Processing*, Modena, Italy, 2007, pp. 9–14.
- [12] Y. Nakata, M. Ando, Lipreading method using color extraction method and eigenspace technique, *Systems and Computers in Japan* 35 (3) (2004) 12–23.
- [13] O. Hua, T. Lee, A new lip feature representation method for video-based bimodal authentication, in: *Proceedings of the 2005 NICTA-HCSNet Multimodal User Interaction Workshop*, vol. 57, Sydney, Australia, pp. 33–37.
- [14] A. Baig, R. Séguier, G. Vaucher, Image sequence analysis using a spatio-temporal coding for automatic lip-reading, in: *Proceedings of IEEE International Conference on Image Analysis and Processing*, Venice, Italy, 1999, pp. 544–549.
- [15] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models—their training and application, *Computer Vision and Image Understanding* 61 (1) (1995) 38–59.
- [16] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (6) (2001) 681–685.
- [17] J. Luetttin, N. Thacker, S. Beet, Speechreading using shape and intensity information, in: *Proceedings of IEEE International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 58–61.
- [18] B. Dalton, R. Kaucic, A. Blake, Automatic speechreading using dynamic contours, in: D. Stork, M. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications*, Berlin, 1996.
- [19] D. Chandramohan, P. Silsbee, A multiple deformable template approach for visual speech recognition, in: *Proceedings of IEEE International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 50–53.
- [20] J. Luetttin, N. Thacker, Speechreading using probabilistic models, *Computer Vision and Image Understanding* 65 (2) (1997) 163–178.
- [21] S. Dupont, J. Luetttin, Audio-visual speech modeling for continuous speech recognition, *IEEE Transactions on Multimedia* 2 (3) (2000) 141–151.
- [22] A. Caplier, Lip detection and tracking, in: *Proceedings of IEEE International Conference on Image Analysis and Processing*, 2001, pp. 8–13.
- [23] I. Matthews, T. Cootes, J. Bangham, Extraction of visual features for lipreading, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (2002) 198–213.
- [24] K. Mase, A. Pentland, Automatic lipreading by optical flow analysis, *Systems and Computers in Japan* 22 (6) (1991) 67–76.
- [25] C. Cedras, M. Shah, Motion-based recognition: a survey, *Image and Vision Computing* 13 (2) (1995) 129–155.
- [26] 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* 15 (2005) 1058–1064.
- [27] D. Xua, X. Li, Z. Liu, Y. Yuan, Cast shadow detection in video segmentation, *Pattern Recognition Letters* 26 (2005) 91–99.
- [28] X. Li, T. Yuan, N. Yu, Y. Yuan, Adaptive color quantization based on perceptive edge protection, *Pattern Recognition Letters* 24 (2003) 3165–3176.
- [29] S. Basu, A. Pentl, A three-dimensional model of human lip motions trained from video, *IEEE Nonrigid and Articulated Motion Workshop*, San Juan, Puerto Rico, 1997, pp. 46–53.
- [30] S. Basu, N. Oliver, A. Pentland, 3D modeling and tracking of human lip motion, in: *Proceedings of IEEE International Conference on Computer Vision*, Bombay, India, 1998, pp. 337–343.