Robust Lip Extraction and Tracking of the Mouth Region

DukSoo Min JinYoung Kim* SeungHo Choi** KiJung Kim***
Chonnnam National University
**Dongshin University
***Kwangyang College
The Department of Electronic Engineering, Chonnam National University
300, Yongbong-dong, Puk-gu, Kwangju, Korea
Tel: +82-62-530-0472, Fax: +82-62-530-0472
E-mail: dsmin@dsp.chonnam.ac.kr, *kimjin@dsp.chonnam.ac.kr

Abstract: Visual features of lip area play an important role in the visual speech information. We are concerned about correct lip area as region of interest (ROI). In this paper, we propose a robust and fast method for locating the mouth corners. Also, we define a region of interest at mouth during speech. A method, which we have used, only uses the horizontal and vertical image operators at mouth area. This searching is performed by fitting the ROI-template to image with illumination control. Most of the lip extraction algorithms are dependent on luminosity of image. We just used the binary image where the variable threshold is applied. The variable threshold varies to illumination condition. In order to control those variations, the gray-tone is converted to binary image by threshold, which is obtained through Multiple Linear Regression Analysis (MLRA) about divided 2D special region.

Thus we obtained the region of interest at mouth area, which is the robust extraction about illumination. A region of interest is automatically extracted.

Keywords: Lip Extraction, Template, Region Of Interest, Regression Analysis

1. Introduction

Robust and accurate visual feature analysis holds a very important role in speech recognition problem. Several researchers have developed the bimodal system, which have demonstrated the valuable tool to improve the robustness of acoustic speech recognizer in noise [10].

The main approaches for extracting visual speech information from image sequences can be grouped into model-based, geometric feature-based and image-based approaches [1,2,3]. In the model-based approach, a model of the visible speech articulators, usually the lip contours, is built and its configuration is described by a small set of parameters. The geometric feature based approach assumes that certain measures such as the height or width of the mouth opening are important features. The image-based approach, the gray-tone image containing the mouth is used directly or after some preprocessing as feature vector [6,7].

In this paper, we propose a robust, fast and cheap scheme for locating the mouth corners. Lip shape information provides part of the visual speech information. Lip parameters are extracted by using automatic extraction algorithm. The lip corners are dependent on the speaker, illumination, reflection, visibility of teeth and mouth opening. Thus feature localization is more difficult than feature tracking. As geometric feature based approach had frequently yielded the error detection in image varieties recently there came alive the trend of the image-based approaches.

We present in this paper an algorithm for lip region extraction robust to different speakers and to variance of illumination. To perform image-based approach, our ROI algorithm is based on a fixed rectangle, which provides another advantage that extra visual information in mouth do not need to be included. A ROI extraction was judged by visual inspection. The results show the superiority of our lip extraction in section 3.

We used a general searching method, which uses an image processing [6,7]. But we used an adaptive statistical method to extract the lip's corners. In mouth area, each of center-divided sides had the dissimilarity of illumination. Specially, there is a marked unbalance when light is lighted from the side. In case of a consistent binary, it will be hard to detect lip corners. In order to control those variations, the gray-tone is converted to binary image by threshold, which is obtained through Multiple Linear Regression Analysis

2. Searching ROI

Subsequent modes describe finer variations, such as lighting direction, reflection, and visibility of teeth and tongue.

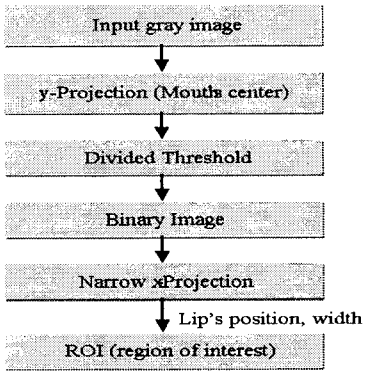


Fig 1. The flowchart of detecting region of interest

We performed the gray-level transformation and equalization, and then binary transformation to extract ROI contained the speaker's mouth. Thus, the image should be conversed to a binary image by a proper threshold value. It is important to find the proper threshold in order to detect the mouth corner. In this paper, we divided an image into four squares and then

computed the adaptive threshold for binary transform on each square with binary coefficients obtained by MLRA.

2.1 detecting lip center

We used a general searching method, which uses an image processing, such as projection. Using the vertical projection, the approximate positions of the lip corners are predicted. To find the vertical position of the line between the lips, we used a vertical integral projection of the gray-tone image in the search region. Since the lip line is the darkest that is horizontally extended structure in the search area, its vertical position can be located where p has its global minimum. We only have to follow, from the center of the image to the left, the minimum row of horizontal projection to find the center of the mouth.

Our first step is to derive rough initial estimates for the coefficients for the semi-parabola corresponding to the top of the upper lip, and the middle of the lower lip. We start by finding the vertical position of the center of the mouth. The vertical position of the center of the mouth can be defined to be the minimum value in each row. In figure 2, all we have to do is follow to the minimum quality of horizontal projection to find lip's vertical position, e.g. from the maximum of the projection to the down.

(a) (b)

Fig 2. The horizontal projection on sample image

2.2 Searching lip corners

The approximate positions of the lip corners are predicted, using the vertical projection. We observe that areas of darkness occurred at the inner position of lips of illumination. The horizontal position of the lips can be found by applying a horizontal binary detector. This area is to the refine search area and regarding the vertical integral projection of this horizontal binary image. The positions of the lip corners can be found by looking for the darkest pixel along the two columns in the search area located at the horizontal boundaries.

In Figure 3, the dotted rectangle is a search region to detect the lip corners. For the binary image, the negative values are dark and the positive values are bright. Unclear image outside the frame of the binary image shows a projection of vertical and horizontal direction. We detected the corners using vertical projection in dotted rectangle.

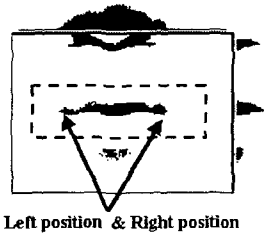


Fig 3. A binary image along with the intensity profile

2.3 Determining the binary threshold

Subsequent modes describe finer variations, such as lighting direction, reflection and visibility of teeth and tongue. In order to control those variations, the gray-tone is converted to binary image by threshold, which is obtained through Multiple Linear Regression Analysis about divided 2D special region.

In mouth area, each of center-divided sides has the dissimilarity of illumination. Specially, there is a marked unbalance when light is lighted from the side. In case of a consistent binary, it will be hard to detect lip corners, like figure 5 upper. To solve this problem we propose that after dividing the image into four sections each binary image should be conversed by different threshold. In detecting a binary threshold appropriate for each binary image, we have investigated about 300 images on our own. The regression analysis was conducted upon whole the binary thresholds that varies to the illumination of each divided image.

$$Th_i = \alpha_i + \beta_{ai} I_a + \beta_{bi} I_b + \beta_{ci} I_c + \beta_{di} I_d + \varepsilon \qquad (1)$$

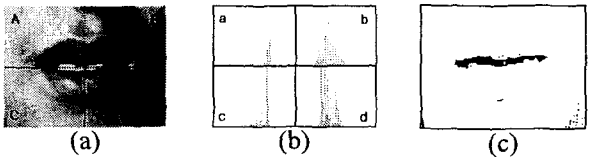
Where I_i is histogram values of each side, Th_i is a suitable threshold to converse a binary image, α and β are a coefficient determined by the regression analysis.

Table 1. Threshold for binary image

	αi	βai	βbi	βci	βdi
Th a	-100.23	0.691	0.441	0.110	0.094
Th b	-59.469	0.286	0.686	0.014	0.127
Th c	-85.124	0.223	0.233	0.528	0.280
Th d	-50.496	0.213	0.226	0.094	0.545

When the values of α and β are derived, the MLRA equation can be written using these values like formula 2. Using the above values endows us with the most conspicuous line suitable for threshold we can use. F-distribution was considered to judge the fitness of presuming dependent variable *Th* on the independent variable *I*. The mean, variance and deviation of the values of independent variables are prerequisite to get F-distribution. Mutual variables have relations each other at meaningful F-ratio, 53.837. To have an interrelation, F-ratio must be generally more than 10 ratios.

Figure 4 shows a binary transformation by the proposed threshold. To confirm the results figure 5 shows a combination image with a binary image and gray-tone image. It was judged by visual inspection about lip corners.



(a) (b) (c)

Fig 4. The illustration of samples to determine the suitable threshold a) a gray-tone image b) Histogram c) binary image (A side histogram is 190, B 133, C 188, D 131 and the binary thresholds are 99, 82, 97 and 79)

Fig 5. A combined image of a binary and gray-tone image

Figure 6 shows the superiority of the binary-transformed image we proposed. The histogram, at (a), of all divided sides is 188, 132, 182 and 133. The (a), (b) and (c) is adjusted the uniform threshold, 126. And the others is adjusted the fixed threshold, we proposed, on all sides. The method we proposed is reported that is suitable to tracking a lip width on a binary image.

To obtain a robust lip's width observation to the lighting condition, we compute the threshold each divided region. Thus, lip's width is detected by local vertical projection in section 2.2.

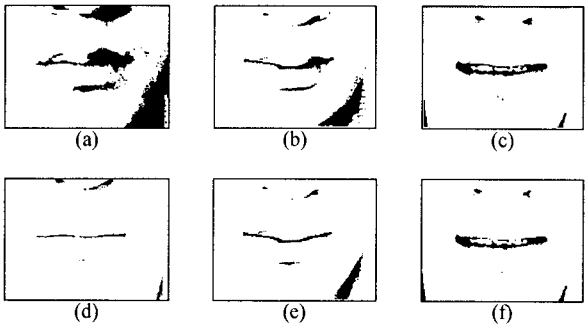


Fig 6. Up samples is the uniform threshold, 126, and down is the suitable threshold of our proposal

3. Experimental Results and ROI Estimation

We tested our method on 1520 MJPEG from the Database consists of gray-tone image sequences of the 22 words, each spoken by 70 male. The male joining in experiment naturally pronounces test word. The frame contains only mouth area and are digitized at 30 frames/sec, 320×240 pixels, 8 bits per pixel.

The ROI search was judged by visual inspection. A region of interest is a portion of an image that we want to process or use to gain some information. A template is a ROI in image and independent from both the speaker and the lighting condition. A template consists of lip's width and lip's height in proportion of 1 to 1. B template is in proportion of 1 to 0.8. Table 2 shows the results for locating the lips produced by using binary image search.

The threshold of searching was classified as ***Good*** if the whole lip was found within template and ***Miss*** otherwise.

Table 2. Results for lip location using binary image search

	Good (%)	Miss (%)
Detection	99.3	0.7
A template	100	0
B template	98.7	1.3

As can be seen from the results so far, the method mentioned above extracting our lips is robust.

In view of the results so far achieved, the method of our lip extraction is robust. Actually it takes comparatively much time to decide a proper binary threshold before a regression analysis is applied because it is conducted wholly by hands. However the lip extraction method is robust in that it is based on extensive sampling, which regards many cases of illumination variances. In addition it is fast in that once a binary threshold is given, the only thing to do is to apply formula 1 to the threshold. It is a robust and fast for locating the mouth corners.

Figure 7 and 8 show the results of ROI detection and lip tracking. When a mouth is highly opened, B template cannot include the whole mouth area.

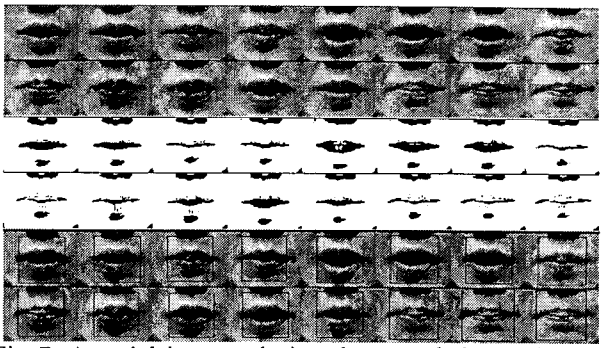


Fig 7. A serial images during the speech intervals. The Upper is an original images, middle is a binary image and down is an images with ROI-template.

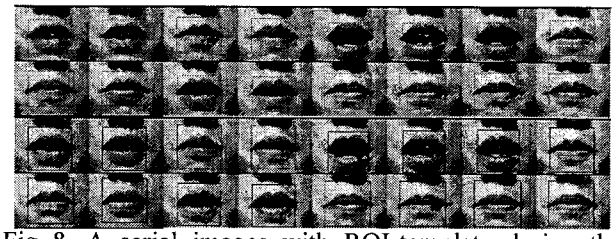


Fig 8. A serial images with ROI-template during the speech intervals. (The upper template is a ratio of 1 to 0.8, and down is a ratio of 1 to 1.)

4. Conclusions

We presented a robust lip extraction, which provides robustness to illumination variance. Our algorithm is simple and fast. This method is highly evaluated in our real-time lipreading since it contains comparatively accurate visual speech information. The extraction, by using a statistical analysis, can be used to track the lip region of the different people and different skin and mouth illumination. Other lip tracking approaches considering intensity variability requires complex theory and costs much.

We will also focus on image sequences processing to enhance the performance of visual speech recognizer. Our main goal is to obtain a robust lipreading performance against image variability.

References

[1]Lievin M. and Luthon F. "Lip features automatic extraction", Proc. Of the 5th IEEE Int. Conf. On Image Processing. Chicago. Illinois, 1998.

[2] J. Luettin, N. A. Tracker, "speechreading using probabilistic models". Computer vision and Image Undestanding, 65(2): 163-178, Feb.1997.

[3] J. Luettin, N. A. Tracker, and S. W. Beet. "Active Shape Models for visual Speech Feature Extraction". Electromic Systems Group Report No.95/44, University of Sheffield, UK, 1995.

[4] A. Yuille, P. Haallinan, and D. S. Cohen. "Feature Extraction from Faces using Deformable templates". International Journal of Computer Vision, 8(2): 99-111, 1992.

[5] P. Radeva and E.marti, "Facial Features Segmentation by Model-based snakes", Int. Conf. On Comp. Anal. And Image Processing, 1995.

[6] Iain Matthews, Tim Cootes. "Lipreading using Shape, Shading and Scale", Auditory Visual Speech Processing, Conf, 1998

[7] Jie Yang, Rainer Stiefelhagen. "Real-Time Face and Facial Feature Tracking and Applications", Auditory Visual Speech Processing, Conf, 1998

[8] B. Moghaddam and A. Pentland. "Probabilistic visual learning for object detection", In IEEE International Conferencd on Computer Vision, pp 786-793, 1995

[9] T. F. Cootes, A. Hill, C. J. Taylor and J. Haslam, "Use of active shape models for locating structures in medical images", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1996

[10] C.Bregler and Yochai Konig, "Eigenlips' for Robust Speech Recognition", Proc. IEEE Int. Conf. On Acoustics, Speech and Signal Processing, pp. 669-672,1994