Lip Localization Technique Towards an Automatic Lip Reading Approach for Myanmar Consonants Recognition

Thein Thein
University of Computer Studies, Mandalay (UCSM)
Mandalay, Myanmar
e-mail: theinthein.cmw@gmail.com

Kalyar Myo San
Faculty of Computer Systems and Technologies (FCST)
University of Computer Studies, Mandalay (UCSM)
Mandalay, Myanmar
e-mail: Kalyar.myosan@gmail.com

***Abstract*—Lip reading system is supportive technology to human being especially for hearing impaired, or elderly people. Lip reading is a process where visual information is extracted by watching lip movements of the speaker with or without sound. So, reliable lip movements are required to extract visual information. To our knowledge, this is the first work for lip movement recognition for Myanmar consonants. So, the major challenge is to recognize lip movements because of many possible lip motions and lip shapes. The accuracy and reliability of speech recognition systems can be improved by using visual information from the movements of the lips, and the need for lip-reading systems continues to grow for every language. Therefore, this paper presents Myanmar consonant recognition based on lip movements towards lip reading by using CIELa*b* color transformation, Moore Neighborhood Tracing Algorithm and linear SVM classifier. The purpose of this study was to develop a visual training technique to accurately identify the characteristics of the lips movement for hearing impairment.**

Keywords-CIELa*b* color space model; lip reading; lip localization; moore neighborhood tracing algorithm; Otsu global thresholding technique

I. INTRODUCTION

Lip reading has been used for various purposes to speech training for the hearing impaired aid and to enhancing speech recognition. In lip reading system, lip localization is the major step to read the lips for extracting visual information from the video input. However, lip movement recognition is active undergoing research topics with lot of improvements that recovers various difficulties faced in the research. Many researchers proposed various techniques to precisely identify the lip area and they presented various features to get significant recognition result. The purpose of this study is to propose a visual teaching method for Myanmar consonants recognition for the hearing impaired persons by precisely localizing the lip movements and activities when they produce the consonants.

In this paper, we propose approaches to recognize accurate lip region based on lip movements for Myanmar consonants recognition. Myanmar consonants can be described in terms of four factors:(1)One syllable consonants, (2)Two syllable consonants, (3)Three syllable consonants, and (4) Four syllable consonant. Here we have tried to extract accurate lip movements for one syllable consonants (င (Nga)၊ ည (Nya)၊ မ (Ma)၊ လ (La)၊ ဝ (Wa)၊ သ (Tha)၊ ဟ(Ha)၊ အ (Ah)) and two syllable consonants (ကျ (Ka Gyi) ၊ ခ (Kha Gway)၊ ဂ (Ga Nge)၊ ဃ (Ga Gyi) ၊ စ (Sa Lone) ၊ ဆ (Sa Lain)၊ ဇ (Za Gwe)၊ ဒ (Da Dway)၊ ဏ (Na Gyi)၊ နှ (Na Nge)၊ ပ (Pa Saug)၊ ဘ (Ba Gone)၊ ရ္ (Ya Gaug)၊ ဠ (La Gyi)) of Myanmar language. This paper aims to extract the accurate upper and lower lip boundary based on lip movement using CIELa*b* color space model, Otsu global thresholding method and Moore Neighborhood Tracing Algorithm.

This paper is organized as follows: the section 1 finished a brief introduction given above, related works will present in section 2, research method will describe in Section 3, experimental results will be mentioned in section 4, conclusion and future work presents in section 5 and finally reference will be done.

II. RELATED WORKS

Many researchers proposed various techniques for lip localization. Namrata Dave implemented a novel color based approach for lip localization based visual feature extraction method which gave a good accuracy for their database [4]. X. Liu and Yiu-ming Cheung proposed a robust lip tracking algorithms using localized color active contours and deformable models [9]. They used a combined semi-ellipse as the initial evolving curve and compute the localized energies in color space to separate from the original lip image into lip and non-lip regions. And then, they presented dynamic radius selection of the local region with a 16-point deformable model to extract the lip. R.E. Hursig et al. developed a robust still image lip localization algorithm designed as a visual front end of a practical AVASR system and presented Gabor filter based facial feature extraction for lip localization [6]. The proposed algorithm is shown to effectively differentiate facial features, including lips, from their backgrounds and to bind the full extent of the lips within a face-classified region of interest and the proposed algorithm making it more versatile within the unconstrained environment. S. Pathan and A. Ghotkar proposed a method

978-1-5386-5384-5/18/$31.00 ©2018 IEEE

to recognize the different English phrases using geometrical features of lip shape [7].

S.S. Morade and B.S. Patnaik presented a novel active contour guided geometrical feature extraction approach [8]. B. Brahme and U. Bhadade presented a methodology for detecting lip using Constrained Local Model (CLM) and extracted geometric features of lip shapes [1]. M. Li and Yiu-ming Cheung proposed automatic lip segmentation approach from a probability model in color space and morphological filter to estimate the model parameter; they used hue and saturation value of each pixel within the lip segment [2]. P. Sujatha et al. presented a new method for automatic lip detection using geometric projection method and adaptive thresholding [5].

III. Research Method

The proposed lip reading system consists of three subsystems. The first is the lip localization system, in which the protrusion in the digital input is placed, and then the feature extraction system that extracts the lip movements suitable for recognizing visual speech. The last one is the classification system. In this study will be carried out to localize upper and lower lip region which are useful for recognize lip movements. Fig. 1 shows the diagram of the detail processing stages for the proposed lip reading system. In this paper, we mainly focus on lip localization to extract accurate lip boundary based on lip movements.

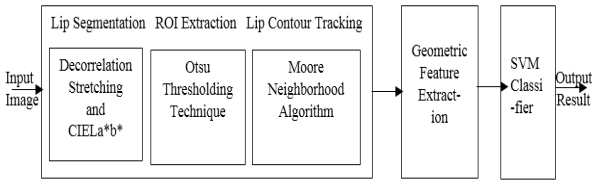


Figure 1. Schematic diagram of the proposed system.

There are three kinds of models for lip reading system. The one is low level or image based, uses mouth region of the image to localize the lips, features of lips and skin pixels are used. It is finding out the height and width of the lips, not the edges of the lips to locate lips. The other one is high level or model based, uses integrity constraints and pixel information to segment the lip. It is finding the corner of the lips to detect the accurate lips. The final one is the hybrid model which is using the parameters of both the models. Among these techniques, we used hybrid approach to for lip reading system.

A. Lip Localization

Lip localization is the important step of lip reading system to detect and extract accurate lip boundary. Lip localization method needs to be performed before the lip features extraction process. Proposed localization stage consists of three steps, namely lip segmentation, lip region extraction, and lip contour tracking to detect lip contour and to localize upper and lower lip boundaries.

Lip Segmentation: Lip segmentation aims to separate the lip from the background skin color. Methods aiming at segmenting the lip shape, boundary or mouth area from the images of the input video. In this paper, Lip region segmentation required several process of image processing. Normalized cross-correlation method, Decorrelation Stretching color enhancing method and CIEL*a*b* color transformation method are used for lip segmentation. The proposed segmentation process starts frame normalization by breaking the video image. Then, extract lip region by using normalize cross-correlation as a preprocessing stage. Fig. 2 shows the original image and the result image after extracting the lip region. Fig. 3 and Fig. 4 show the normalized image frames. The image lip information is in RGB color space. So, to subtract the robust lip region, RGB color scheme of the image is not improper for immediate processing because it contains a lot of mixed information about lightness etc. Decorrelation Stretching color enhancing method and CIEL*a*b* color transformation method are based on differences in color composition between lip as the object and skin as the background.

(a) (b)

Figure 2. (a) Original image (b) Segmented lip region.

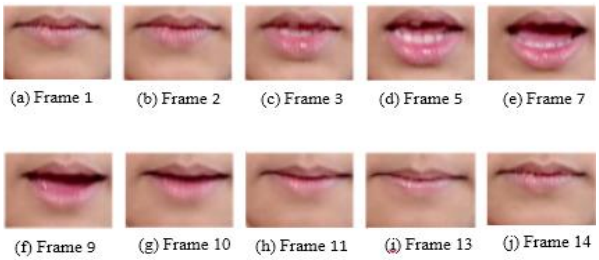


Figure 3. (a) to (j) Number of selected frames for utterance of Nya (one syllable consonant).

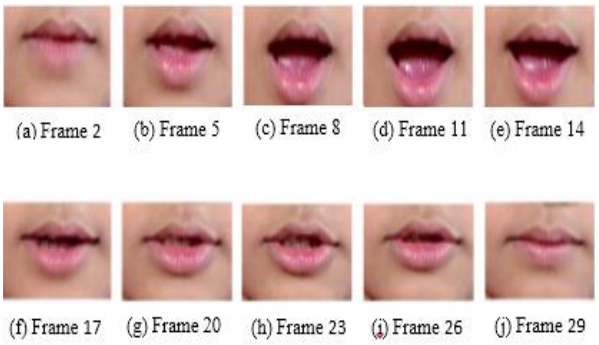
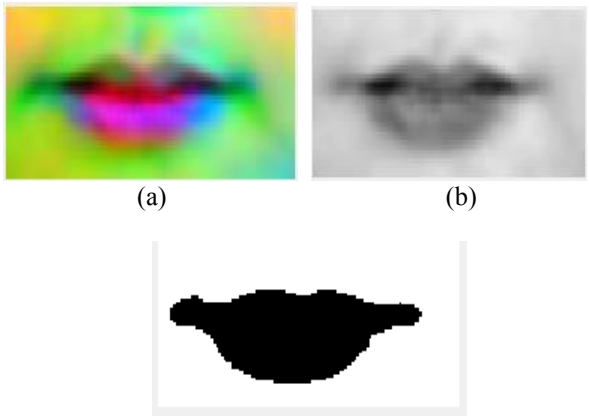


Figure 4. (a) to (j) Number of selected frames for utterance of Ga Gyi (two syllable consonant).



(a) (b)

(c)

Figure 5. (a) Color enhanced image, (b) Color transformed image , (c) Equalized image.

Skin colors are marked more on color composition compare to brightness, even on different people. Color compositions of skins are remarkably constant even when exposed by a lot of illumination. So, in our experiments, we used Decorrelation Stretching color enhancing method with Stretch limit. And then, we used Kalman image filtering method to eliminate a lot of noise and unwanted information in image around mouth region. Fig. 5(a) showed the color enhanced image.

After enhancing image, RGB color image is transform into CIEL*a*b* color space based on first layer L channel, Fig. 5(b) showed the result of enhanced image. For extract lip region exactly and accurately, histogram equalization is used to color contrast. Therefore, Fig. 5(c) showed that the lip area appears much darker than the skin.

1) ROI extraction: Otsu thresholding is an automatic thresholding technique that commonly referred to as adaptive threshold [3]. Otsu thresholding is needed to perform image binarization to the image resulted from color transformation. This Otsu thresholding method calculates the value of threshold (T) for segmentation based on the input image. Otsu thresholding technique seeks the optimal threshold value to separate object from background by maximizing the variance between classes while minimizing the variance within classes. The maximum value of the variance between classes is defined in the following equations:

$$\sigma^2_w(T) = \max_t \sigma^2_w(t)$$
$$\sigma^2_w(t) = w_1(t)\, w_2(t)\, (\mu_2(t) - \mu_1(t)) \quad (1)$$

where w_1 and w_2 are the probability of pixels in each class, while μ_1 and μ_2 are the mean gray scale of each class. The probabilities and means for each class are updated iteratively. In this paper, the upper and lower lip contour is extracted by using Otsu global thresholding technique; Fig. 6 shows the result of lip boundary contour.

Figure 6. Extracted lip boundary contour.

2) Lip contour tracking

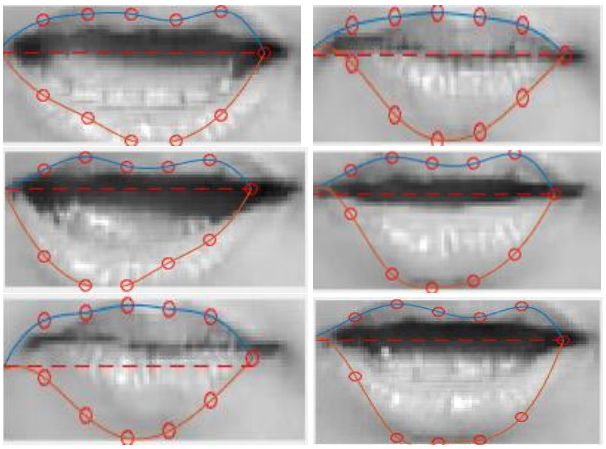


Figure 7. Lip tracking results for utterance of Ga Gyi (two syllable consonant) on only selected frame. The first column shows negative tracking results and the second column shows positive tracking results.

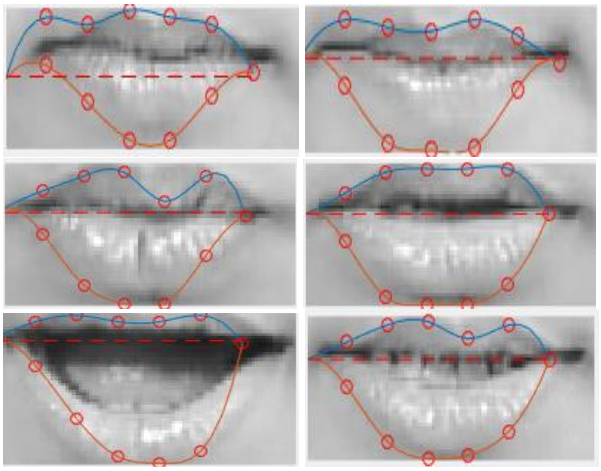


Figure 8. Lip tracking results for utterance of Nya (one syllable consonant) on only selected frame. The first column shows negative tracking results and the second column shows positive tracking results.

Tracing and localizing for lip boundary process is difficult and not always efficient. Therefore, this paper proposes lips tracing algorithm. There are four of the most common contour tracing algorithms, namely: the Square Tracing algorithm, Moore-Neighborhood Tracing Algorithm, Radial Sweep and Theo Pavlidis' Algorithm. The first two are easy to implement and are therefore used frequently to trace the contour of a given pattern. In this paper, after extracting the lip contour of the previous lip frame, we let it be the initial evolving curve embedded in Moore

Neighborhood tracing algorithm to localize lip boundary and to lip contour tracking of the current frame. With this approach, we are able to estimate lip boundary in each frame of video sequence. Fig. 7 and Fig. 8 showed an example of lip tracking results.

IV. Experimental Results

A. Database of Lip Reading System

In this paper, experiments were tested on own audiovisual database of Myanmar consonants. Database consists of twelve speakers, four persons of male speakers and eight persons of female speakers. Both are white and black skin. Sony DVCan-DSR 300A professional video camera is used. Videos are recorded in mp4 format with 23frames/second. Recording distance is constant. Each image frame has resolution of 720×480.

B. Results of Lip Localization

TABLE I. Localization Accuracy Rate for Two Syllable Consonants

Two Syllable Consonants	Accuracy Rate		Error Rate	
	YCbCr	*CIELa*b**	*YCbCr*	*CIELa*b**
ကျ (Ka Gyi)	91.66%	93.30%	8.34%	6.70%
ခ (Kha Gway)	86.79%	90.57%	13.21%	9.43%
ဂ (Ga Nge)	91.66%	91.66%	8.34%	8.34%
ဃ (Ga Gyi)	79.62%	90.74%	20.38%	9.26%
စ (Sa Lone)	86.00%	90.00%	14.00%	10.00%
ဆ (Sa Lain)	89.00%	94.55%	11.00%	5.54%
ဇ (Za Gwe)	86.66%	91.80%	13.34%	8.20%
ဏ (Na Gyi)	86.66%	86.66%	13.34%	13.34%
ဒ (Da Dway)	86.20%	91.38%	13.80%	8.62%
နှ (Na Nge)	86.00%	90.00%	14.00%	10.00%
ပ (Pa Saug)	89.09%	92.73%	10.91%	7.27%
ဘ (Ba Gone)	87.75%	91.84%	12.25%	8.16%
ရ့ (Ya Gaug)	84.44%	88.89%	15.56%	11.11%
ဠ (La Gyi)	88.13%	93.22%	11.87%	6.78%
Total accuracy rate/ error rate	87.11%	91.23%	12.89%	8.77%

TABLE II. Localization Accuracy Rate for One Syllable Consonants

One Syllable Consonants	Localization Accuracy Rate		Error Rate	
	YCbCr	*CIELa*b**	*YCbCr*	*CIELa*b**
င(Nga)	87.87%	90.9%	12.13%	9.10%
ည(Nya)	78.57%	88.09%	21.43%	11.91%
မ(Ma)	94.44%	97.2%	5.56%	2.80%
လ(La)	83.78%	89.19%	16.22%	10.81%
ဝ(Wa)	81.81%	89.09%	18.19%	10.91%
သ(Tha)	93.18%	93.18%	6.82%	6.82%
ဟ(Ha)	89.47%	89.47%	10.53%	10.53%
အ(Ah)	88.67%	92.45%	11.33%	7.55%
Total accuracy rate/ error rate	87.22%	91.20%	12.78%	8.80%

The objective of this paper is to detect, localize and track the lip movement to reach the goal of an automatic and robust lip reading. Significant localization accuracy is needed to get accurate and precise recognition results for lip reading system. So, we experimented the segmentation process with YCbCr and CIELa*b* color space model. Table (1) and Table (2) show the localization accuracy rate varies on two color space model. These results demonstrated that when we used CIELa*b* color space in segmentation process, the localization accuracy rate increase. The presented lip localization techniques have achieved a more satisfactory result for recognition.

V. Conclusion and Future Work

This paper proposed the technique to localize lip region for the Myanmar consonants recognition. The experimental system demonstrates that this technique performs lip motion sequences in video. In our experiment, we can localize all of the test lip movement successfully and the results were perceived to be acceptable for lip reading. For future work, we will intend to explore more observable features and recognition phase by applying Support Vector Machine classifier (SVM) method for remaining three syllable and four syllable Myanmar consonants recognition based on lip movements to produce the good recognition result. We hope that this study will help to a new teaching and learning method for Myanmar language education.

References

[1] Brahme and U. Bhadade, "Lip detection and lip geometric feature extraction using constrained local model for spoken language identification using visual speech recognition", Indian Journal Science and Technology, Vol 9(32), August 2016.

[2] M. Li and Y.M. Cheung, "Automatic segmentation of color lip images based on Morphological filter", *ICANN*, 2010.

[3] N. Otsu, "A Threshold selection method from Gray-Level Histogram", IEEE Transaction On Systems, Man, and Cybernetics, Vol. SMC-9, Pontificia Universidance Catolica Do Rio De Janeiro, 1979.

[4] Namrata Dave, "A Lip localization based visual feature extraction method", Electrical & Computer Engineering: An International Journal, ECIJ, Volume 4, Number 4, 2015.

[5] P.Sujatha et al., "Novel pixel-based approach for mouth localization", International Journal of Computer Applications, (0975 – 8887) , International Conference on Computing and information Technology, *IC2IT*, 2013.

[6] R.E. Hursig, J.X. Zhang, and C. Kam, "Lip localization algorithm using Gabor filters", International Conference on Image Processing, Computer Vision and Pattern Recognition ,ICPR, 2012.

[7] S. Pathan and A. Ghotkar, “Recognition of spoken English phrases using visual features extraction and classification”, International Journal of Computer Science and Information Technologies, IJCSIT, Vol.6 (4), 3716-3719, 2015.

[8] S.S. Morade and B.S. Patnaik, “Automatic lip tracking and extraction of lip geometric features for lip reading”, International Journal of Machine Learning and Computing, Vol 3, No.2, April 2013.

[9] X. Liu and Y.M. Cheung, “A Robust lip tracking algorithm using localized color active contour and deformable models”, *ICASSP*, 2011