รายการอ้างอิง

1. J.S. Kim, W. Jang and Z. Bien, A Dynamic Gesture Recognition System for The Korean Sign Language (KSL) Systems IEEE Transactions on Man and Cybernetics (April 1996) :354 – 359
2. K. Imagawa, S. Lu and S. Igi, Color-Based Hands Tracking System for Sign Language Recognition Face and Gesture (1998): 462–467.
3. N. Tanibata, N. Shimada and Y. Shirai, Extraction of Hand Features For Recognition of Sign Language Words Proc. of Int. Conf. on Vision Interface (2002): 391-398.
4. S. Lu, S. Igi, H. Matsuo and Y. Nagashima, Towards a Dialogue System Based on Recognition and Synthesis of Japanese Sign Language. Proc. Beifield Gesture Workshop 1997 (1997).
5. T. Starner, J. Weaver and A. Pentland, Real-time American Sign Language Recognition Using Desk and Wearable Computer Based Video IEEE Transactions on Information Technology in Biomedicine (December 1998) : 1371 – 1375.
6. J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink and G. Sagerer, Improving Adaptive Skin Color Segmentation by Incorporating Results from Face Detection IEEE Int. Workshop on Robot and Human Interactive Communication (September 2002).
7. J. Yang and A. Waibel, A Real-Time Face Tracker Proc. of 3rd Workshop on Applications of Computer Vision (1996): 142-147.
8. Q. Huynh-Thu, M. Meguro, M. Kaneko, Skin-Color Extraction in Images with Complex Background and Varying Illumination 6th IEEE Workshop on Applications Computer Vision (December 2002).
9. C. Garcia and G. Tziritas, Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis IEEE Transactions on Multimedia (September 1999): 264-277.

10. D. Chai and K.N. Ngan, Face Segmentation Using Skin Color Map in Videophone Applications IEEE Transactions Circuits and Systems for Video Technology (1999): 551-564.

11. N. Habili, C.C. Lim and A. Moini, Segmentation of the Face and Hands in Sign Language Video Sequences Using Color and Motion Cues IEEE Transactions on Circuits and Systems for Video Technology (August 2004).

12. R.L. Hsu, M. Abdel-Mottaleb, A.K. Jain, Face Detection in Color Images. IEEE Transactions on Pattern Analysis and Machine Intelligence (May 2002): 696-706.

13. S.L. Phung, A. Bouzerdoum and D. Chai, A Novel Skin Color Model in YCbCr Color Space and Its Application To Human Face Detection IEEE Int'l Conf. on Image Processing (September 2002): 289-292.

14. A. Hadid, M. Pietikainen and B. MartinKauppi, Color-Based Face Detection Using Skin Locus Model Hierarchical Filtering Proc. IEEE Pattern Recognition (2002): 196-200.

15. A. Shamaie and A. Sutherland, A Dynamic Model for Real-time Tracking of Hands in Bimanual Movements 5th International Workshop on Gesture and Sign Language based Human-Computer Interaction (April 2003): 15-17.

16. S. Kawato, and J. Ohya, Automatic Skin-color Distribution Extraction for Face Detection and Tracking. ICSP2000: The 5th International Conference on Signal Processing (August 2000): 1415-1418.

17. B.D. Zarit, B.J. Super, and F. Quek, Comparison of Five Color Models in Skin Pixel Classification Proc. Int'l Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (September 1999): 58-63.

18. J.-C. Terrillon, M.N. Shirazi, H. Fukamachi, and S. Akamatsu, Comparative Performance of Different Skin Chrominance Models and Chrominance Spaces for the Automatic Detection of Human Faces in Color Images Proc. IEEE Int'l Conf. on Face and Gesture Recognition (2000): 54-61.

19. D.M. Saxe and R.A. Foulds, Robust Region of Interest Coding for Improved Sign Language Telecommunication IEEE Transactions on Information Technology in Biomedicine (December 2002): 310-316.

20. M.J. Jones and J.M. Rehg, Statistical Color Models with Application to Skin Detection In Proc. of the Computer Vision and Pattern Recognition 1999 (1999): 274–280.

21. วีดิทัศน์ประกอบการเรียนภาษามือไทย กรมสามัญศึกษา กระทรวงศึกษาธิการ จัดทำโดย สมาคมคนหูหนวกแห่งประเทศไทย

22. M.C. Shin, K.I. Chang and L.V. Tsap, Does Colorspace Transformation Make Any Difference on Skin Color IEEE Workshop on Applications of Computer Vision (2002): 275-279.

23. V. Vezhnevets, V. Sazonov and A. Andreeva, A Survey on Pixel-Based Skin Color Detection Techniques Proc. Graphicon (2003): 85-92.

29. U.M. Erdem, and S. Sclaroff, Automatic Detection of Relevant Head Gestures in American Sign Language Communication Proc. International Conference on Pattern Recognition 2002 (August 2002).

24. C.R. Wren, A. Azarbayejani, T. Darrel, and A.P. Pentland, Pfinder: Real-Time Tracking of the Human Body IEEE Transaction on Pattern Analysis and Machine Intelligence (July 1997): 780-785.

25. I. Haritaoglu, D. Harwood, and L.S. Davis, W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People Proc. Computer Vision and Pattern Recognition (2000).

26. L.G. Shapiro and G.C. Stockman, Computer Vision Publisher Prentice Hall, 2001.

27. M. Sonka, V. Hlavac, and R. Boyle, Image Processing Analysis and Machine Vision 2nd Ed., Brooks/Cole Publishing Company, 1999.

28. R.C. Gonzales and R. E. Woods, Digital Image Processing Addision Wesley, 1992.

30. Z. M. Hafed, Object tracking Principles and challenges to keeping on eye on things IEEE Potentials, August/September 1999.

ภาคผนวก

บทความทางวิชาการที่ได้รับการเผยแพร่

1. N. Soontranon, S. Aramvith and T.H. Chalidabhongse "Face and Hands Localization and Tracking for Sign Language Recognition," Proceeding of International Symposium on Communications and Information Technologies (ISCIT 2004), October 2004, Saupaulo, Japan
2. นรุตม์ สุนทรานนท์, สุภาวดี อร่ามวิทย์ และ ธนารัตน์ ชลิดาพงศ์, "การตรวจหาและติดตามใบหน้าและมือสำหรับวีดิทัศน์ภาษามือ," การประชุมทางวิชาการวิศวกรรมไฟฟ้า (EECON-27) เล่มที่ 2 หน้า 157-160, พฤศจิกายน 2547, ขอนแก่น, ประเทศไทย
3. N. Soontranon, S. Aramvith and T.H. Chalidabhongse "Improved Face and Hand Tracking for Sign Language Recognition," To appear in International Conference on Information Technology (ITCC 2005), April 2005, Lasvegas, USA

Face and Hands Localization and Tracking for Sign Language Recognition

N. Soontranon[1], S. Aramvith[1] and T. H. Chalidabhongse[2]
[1]Department of Electrical Engineering
Chulalongkorn University
Bangkok 10330 Thailand
Tel: +66-2218-6909
E-mail: Supavadee.A@chula.ac.th
[2]Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520 Thailand
Tel: +66-2737-2551 Ext.526
E-mail: thanarat@it.kmitl.ac.th

Abstract: In this paper, we develop the face and hand detection and tracking for sign language recognition system. We first perform preliminary evaluation on several color spaces to find the most suitable one using non-parametric model approach. Then, we propose to use the elliptical model on CbCr to lower the complexity of the detection algorithm and to better model the skin color. After the skin regions from the input video have been segmented, the interested facial features and hands are detected using luminance differences and skeleton features respectively. In the tracking stage, each blob determines search region and find MMSE (Minimum Mean Square Error) to match its own blob. The block matching method between previous and current frame is used. Experimental results show that our proposed system is able to detect and tracking face and hands of sign language video sequences.

1. Introduction

Rapid advances in information and communication technologies (ICTs) are central to transformations in the way people conduct in their everyday lives. Information and knowledge are expanding in quantity and accessibility. However, people with functional limitations, such as deaf people, often experience wide communication gaps. Due to the hearing limitation, deaf people have developed their own culture and methods for communicating among them as well as with hearing groups by relying on signing. The sign language translator is thus an important tool to enabling effective communication between deaf and hearing people. As of now, there have been several researches on sign language recognition system using vision based approach [3,4]. These systems track hands during sign language to recognize the movement, but some of them can support only single-handed sign language or the situation where the signer wears long-sleeve shirt. Thus, it cannot capture real sign language motion.

In this work, we develop the face and hand detection and tracking for sign language recognition system. We construct non-parametric model of the skin color using CbCr color space. Then, we propose to use the elliptical model on CbCr to lower the complexity of the detection algorithm and to better model the skin color. After the skin regions from the input video have been segmented, the interested facial features and hands are detected using luminance differences and skeleton features respectively.

The proposed framework consists of two stages: initial and tracking stages. In the initial stage, the face, arm and/or hand will be detected. This information will be used as an input in the tracking stage. The system is tested using Thai sign language video sequences [9].

The organization of this paper is as follows. In section 2, we describe three color spaces, YCbCr, normalized RGB, and HSV and choose a skin-color model in CbCr color space. Section 3 describes the features detection: facial features are extracted by using the difference of luminance component between skin and facial features, while arm or hand is detected by skeleton method. Section 4, we present the proposed framework. Section 5, we show the experimental results. Section 6 concludes the paper.

2. Skin-color Segmentation

To locate and track the signer's face and hands, we first segment the regions from the input video using human skin-color model. The skin-color detection confuses under different lighting conditions while the signer is moving his face and hands, thus we need to use a proper color space.

2.1 Color Space

To date, various color models have been proposed for skin-color detection, e.g., the CbCr color space is used in [1,6], the rg color space is used in [2], the HS color space is used in [3], and etc. Considering the color spaces that are similar to HVS (Human Vision System), we selected to evaluate three color spaces that separate luminance from the chrominance, YCbCr, normalized RGB, and HSV.

YCbCr

The YCbCr was defined by the BT.601 [5] recommendation as a digital color coordinate, and it has been used in various video compression standards such as H.26L, H.264, MPEG-4 and etc.

The YCbCr values are related to the RGB values by:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.2568 & 0.5041 & 0.0980 \\ -0.1482 & -0.2910 & 0.4392 \\ 0.4392 & -0.3678 & -0.0714 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \quad (1)$$

The inverse transformation is:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1.1643 & -0.0018 & 1.5958 \\ 1.1643 & -0.3914 & -0.8135 \\ 1.1643 & 2.0178 & -0.0012 \end{bmatrix} \begin{bmatrix} Y-16 \\ Cb-128 \\ Cr-128 \end{bmatrix} \quad (2)$$

In YCbCr color space, Y reflects the luminance and is scaled to have a range of (16-235); Cb and Cr are scaled versions of color differences B-Y and R-Y, respectively. The scaling and shifting are designed so that they have a range of (16-240).

Normalized RGB (rgb)

Another color space which considers only pure color is normalized RGB, and is denoted as "rgb":

$$r = \frac{R}{R+G+B} \quad (3)$$

$$g = \frac{G}{R+G+B} \quad (4)$$

$$b = \frac{B}{R+G+B} \quad (5)$$

where (r,g,b) are normalized color values of (R,G,B) respectively. Chromatic color (r,g) is known as pure colors in the absence of brightness. In fact (r,g) define $R^3 \Rightarrow R^2$ mapping. Color blue is redundant after normalized because r+g+b = 1. The range of (r,g) is from 0 to 1.

HSV

HSV (Hue, Saturation, Value) color space is used in many algorithms that segment skin-color regions. It is also compatible with human vision system. The HSV color space is obtained by non-linear transformation of the RGB color space.

$$H = \begin{cases} H_1 & ;B \le G \\ 360^\circ - H_1 & ;B > G \end{cases} \quad (6)$$

where

$$H_1 = \cos^{-1}\left\{\frac{0.5[(R-G)+(R-B)]}{\sqrt{(R-G)(R-G)+(R-B)(G-B)}}\right\} \quad (7)$$

$$S = \frac{\max(R,G,B) - \min(R,G,B)}{\max(R,G,B)} \quad (8)$$

$$V = \frac{\max(R,G,B)}{255} \quad (9)$$

In HSV, H and S are chrominance components while V is luminance component. For skin-color detection that is robust to illumination change, we ignore the luminance (V), instead, we consider only chrominance H and S. The range of H is from 0 to 360 degree and S is from 0 to 1.

In this research, the non-parametric skin color modeling is used to compare the accuracy of skin-color detections using these three color-spaces. The results (shown in Section 5) suggested us to use the YcbCr as a color space for skin color modeling.

2.2 Elliptical CbCr Skin-color Model

In order to make the system robust to the variable lighting condition, we use only the chrominance components CbCr-in detecting face and hands. Figure 1(a) illustrates the distribution of skin color in the 3 dimensional YCbCr space. Considering only the chrominance components in Figure 1(b), we found that the color of human skin formed up an elliptical shape. Thus, the human skin color cluster can be modeled with an ellipse:

$$\frac{L_1^2}{a^2} + \frac{L_2^2}{b^2} = 1 \quad (10)$$

where a is the size of ellipse at major axis, b is the size of ellipse at minor axis, L_1 and L_2 are the major and minor axis of the ellipse respectively and obtained by:

$$L_1 : AC_b + BC_r + C = 0 \quad (11)$$

$$L_2 : BC_b - AC_r + D = 0 \quad (12)$$

where A, B, C and D can be obtained from mean and covariance of (C_b, C_r), in which L_1 and L_2 have intersection at mean (C_b, C_r) and the covariance (C_b, C_r) is used a slope of the L_1 and L_2.

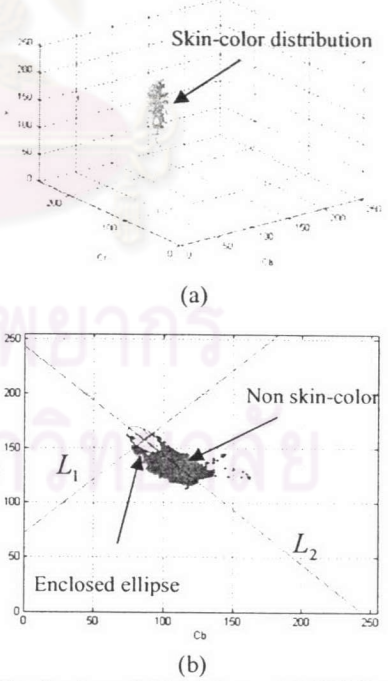


Figure 1. Distribution of skin color on (a) YCbCr (b) CbCr.

In detection, a pixel (i, j) is marked as a skin color pixel if it is inside the learned elliptical model in CbCr subspace:

$$MSkin(i,j) = \begin{cases} 1, & \text{if enclose with ellipse} \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

$MSkin(i, j)$ is the skin-color mask where i and j are row and column of the pixel in the input image. Figure 2 illustrates the result of skin color segmentation using elliptical CbCr skin color model.

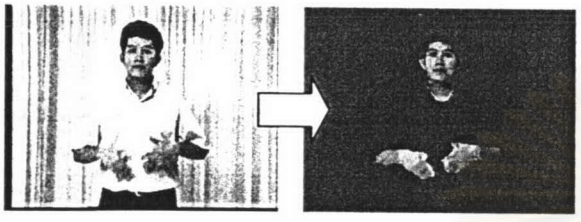
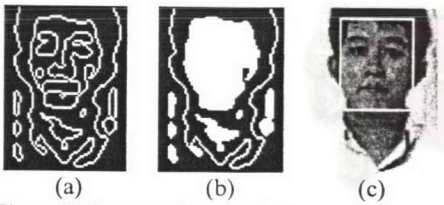


Figure 2. Skin color segmentation using elliptical CbCr skin color model.

3. Facial Features and Hands Localization

In order to recognize the sign language, the positions and motions of salient features such as hands, head, facial features must be analyzed and interpreted. In this research, after segmenting the skin color regions from the input video, the salient facial features and hands are then located and tracked.

3.1 Facial Features



(a) (b) (c)

Figure 3. Separate face region
(a) Edge detection.
(b) Fills the holes in the binary image (a).
(c) Face region.

- Separating face region from the neck

Because we work on sign language videos, the signers are always in upright position. This allows us to assume the upper detected skin color blob is the head of the signer that is composed of signer's face and neck skin. To accurately compare with the reference face template, we need to separate only face region from the head area. Our method uses edge detection with zero-cross, then fills the holes in enclosed boundary. The face region is defined as the biggest component after performing the connected component analysis (Figure 3).

- Detecting eyes, nose and mouth

After separating the face region from the neck, the salient facial features such as eyes, nose, and mouth are then detected (Figure 4). These facial features contain stronger luminance component than other facial region. Thus, we segment the interested facial features by thresholding the facial image as follows:

$$TH1 = Mean(Y(i_f, j_f)) - \beta * Std(Y(i_f, j_f)) \quad (14)$$

where $Y(i_f, j_f)$: luminance of face region pixels.

β : weighting factor equal 1.5.

$$MFacial(i_f, j_f) = \begin{cases} 0, if\ MSkin(i_f, j_f) \le TH1 \\ 1, if\ MSkin(i_f, j_f) > TH1 \end{cases} \quad (15)$$

$MFacial(i_f, j_f)$: facial features mask when i_f and j_f are group of face region pixels.

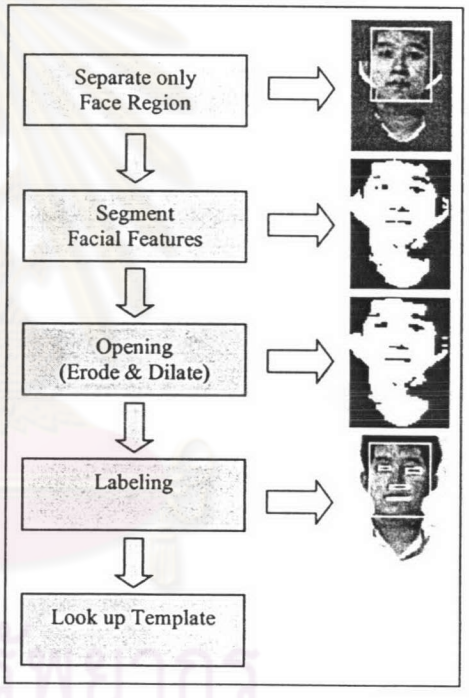


Figure 4. Facial features detection algorithms.

3.2 Arm and Hand

When the signer wearing long-sleeve shirt, the hand locations are easily computed. In contrast, in the case of the signer wearing short-sleeve shirt, it is not quite easy to locate and track the hands. We need to separate the hand from the arm region. To do this, the hand and arm region is thinned to obtain the skeleton of the region. The hand is then defined as the dense intersections region (Figure 5).

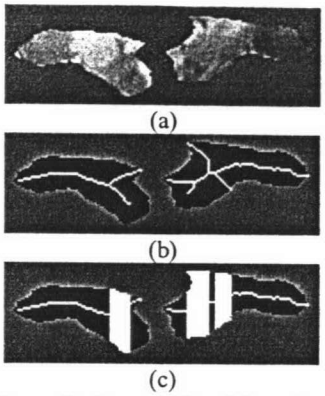


Figure 5. Arm and hand detection.
(a) Detected skin blobs.
(b) After applying thinning.
(c) Hand regions are marked.

4. System Framework

The system is divided into 2 stages; the initial and tracking stages. In initial stage (Figure 6 (a)), the RGB sequence is transformed to YCbCr. The elliptical CbCr skin color model is constructed and used to segment skin-color regions of each video frame. The segmented skin-color regions are closed (dilation and erosion) to fill up the small holes. Next, the connected component analysis is applied to eliminate some noise regions and to localize the salient skin color blobs. Then, the facial features and hands localization explained in Section 3 is performed to initially specify the positions of those interested features needed in further tracking. In tracking stage (Figure 6(b)), each blob determines search region and find MMSE (Minimum Mean Square Error) to match its own blob. The block matching method between previous and current frame is used:

$$(X,Y) = \arg \min_{dx,dy \in S} Sum(DFD) \qquad (16)$$

$$DFD = \frac{1}{N}\{I_{t-1}(x,y) - I_t(x+dx, y+dy)\}^2 \qquad (17)$$

where DFD : Displaced Frame Difference.
I_{t-1} : previous frame.
I_t : current frame.
dx, dy : search region.
N : block size.

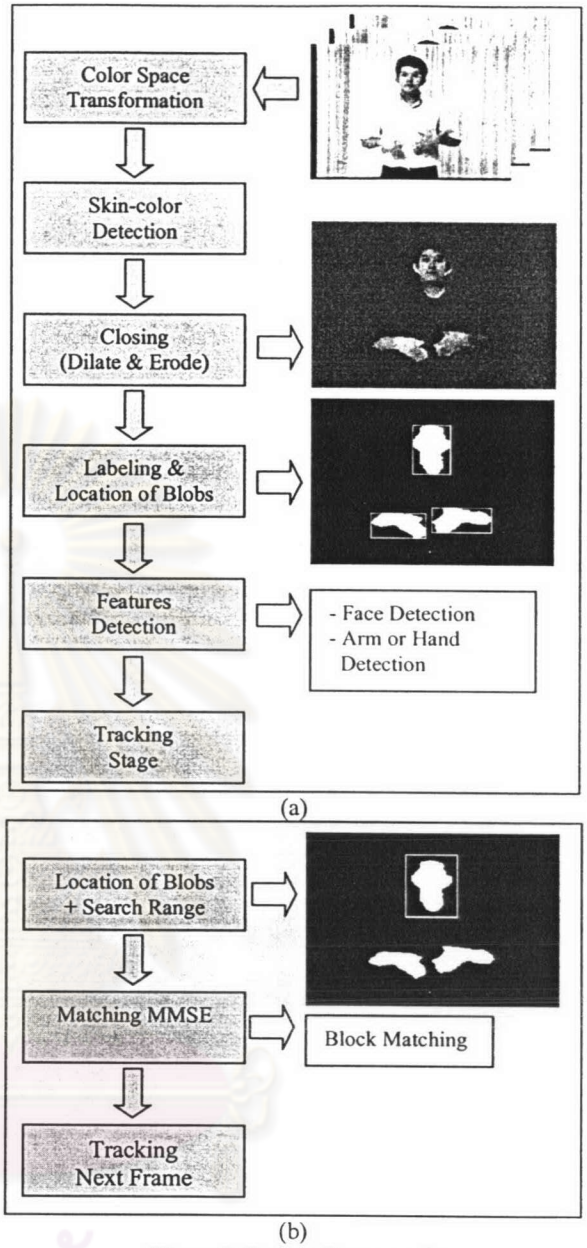


Figure 6. System Framework
(a) Initial stage
(b) Tracking stage

5. Experimental Results

In this section, we demonstrate the results of our proposed algorithm on four video sequences of Thai Sign Language signing shown in Figure 7. All of them are in different lighting conditions, and the signers wear both long-sleeve and short-sleeve shirts.

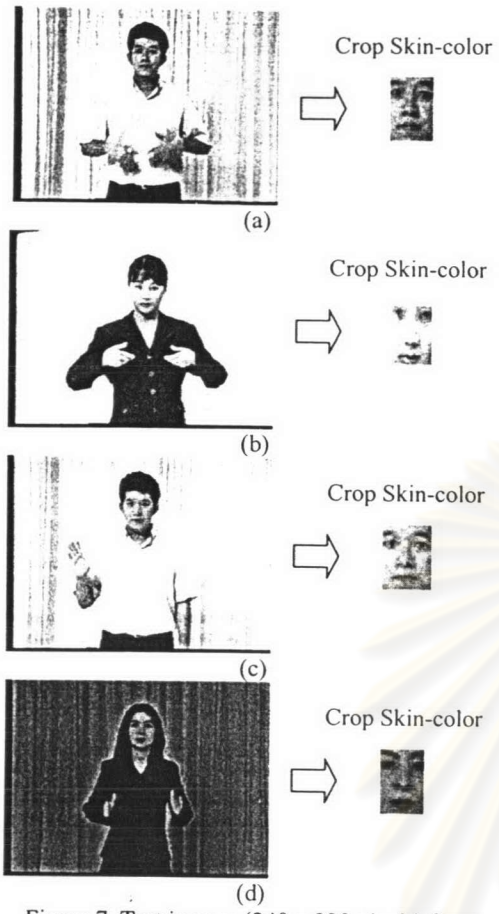


Figure 7. Test images (240 x 320 pixels) from Thai sign language video sequences [9].

The first experiment is the comparative evaluation of skin color modeling and detection using three candidate color spaces -YCbCr, Normalized RGB, and HSV-. The performance of skin color modeling and detection can be measured by two parameters: detection rate (DR) and false alarm rate (FAR), which represent in equation (18) and (19) respectively. The comparative results are shown in Table 1.

$$DR = \frac{TP}{TP + FN} \times 100\% \quad (18)$$

$$FAR = \frac{FP}{TP + FP} \times 100\% \quad (19)$$

where TP: True Positive
FP: False Positive
FN: False Negative

Table 1 shows superior results of CbCr skin color model over the other two for the test videos. This suggests us to go along with the YCbCr color metric for skin color modeling and detection in this research. As described in Section 2, we further model the skin color using elliptical shape on CbCr space. Table 2 shows the overall quantitative detection performance on the four test videos.

For qualitative performance, the results are shown in Figure 8.

Table 1 Comparative performance of skin color detection in each color space using non-parametric skin color model.

Image a

Color Space	DR(%)	FAR(%)
CbCr	68.85	1.67
rg	64.55	6.45
HS	52.89	8.54

Image b

Color Space	DR(%)	FAR(%)
CbCr	59.62	2.65
rg	64.72	70.54
HS	48.75	68.74

Image c

Color Space	DR(%)	FAR(%)
CbCr	67.91	1.66
rg	73.08	6.13
HS	58.86	4.29

Image d

Color Space	DR(%)	FAR(%)
CbCr	68.27	0.95
rg	59.45	81.86
HS	55.45	83. 24

Table 2 Performance of skin color detection using elliptical CbCr model on each video sequence.

Image	DR(%)	FAR(%)
A	91.15	4.27
B	79.60	4.15
C	92.19	6.97
D	96.26	7.94

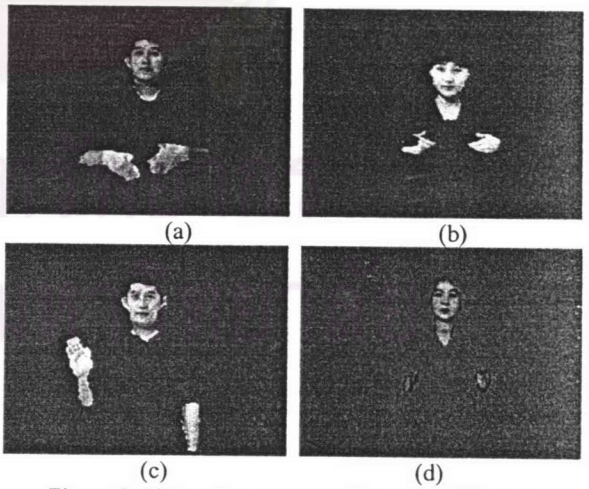
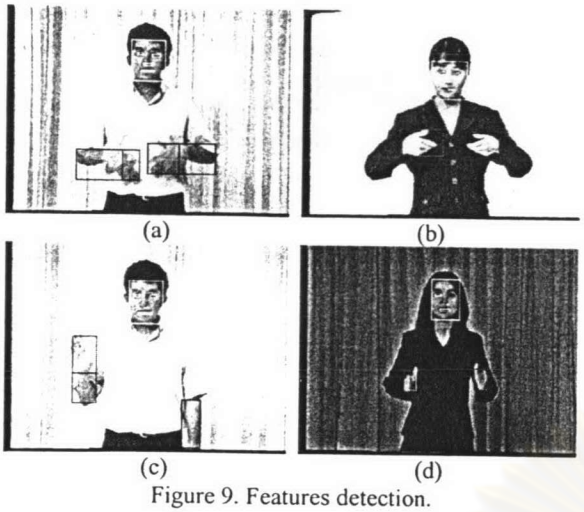
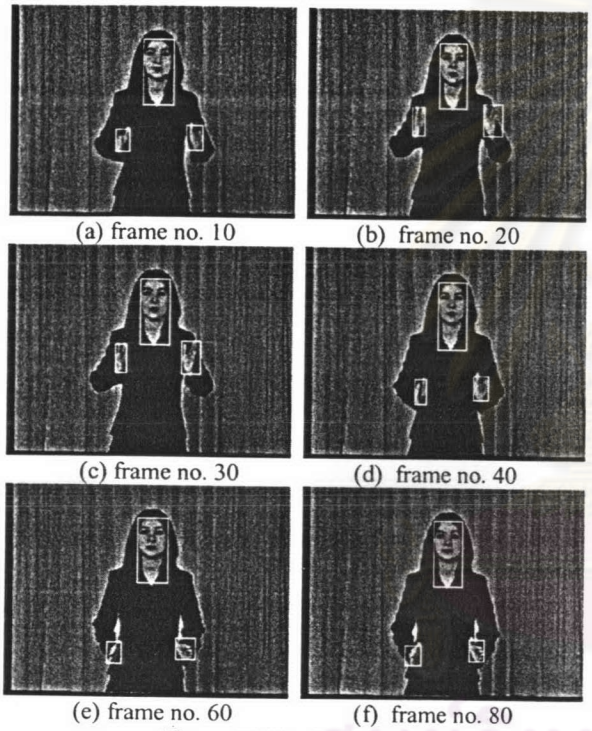


Figure 8. Skin color segmentation using elliptical CbCr model on each video sequence



(a) (b) (c) (d)

Figure 9. Features detection.



(a) frame no. 10 (b) frame no. 20 (c) frame no. 30 (d) frame no. 40 (e) frame no. 60 (f) frame no. 80

Figure 10. Tracking result.

6. Conclusions and Future Work

We have presented the face and hand detection and tracking for sign language recognition system. First, we compared skin color modeling and detection using three color-spaces by constructing non-parametric model of the skin color. The CbCr yields superior result over the other two. To lower the complexity of the detection algorithm and to better model the skin color, the elliptical model on CbCr is used. After segmenting the skin regions from the input video, the interested facial features and hands are detected using luminance differences and skeleton features respectively.

Currently, we are working on temporal information analysis to improve the algorithm in locating and tracking the face and hands. Temporal information will be used by compute distance of moving face and hands under assumption that face changes position less than hands in this signing circumstance. We are also working on solving the region merging and splitting problems that happen when the signer moves his hands over his face, cross each other, etc.

Acknowledgment

The authors are very grateful Cooperative Project between Department of Electrical Engineering and Private Sector for Research and Development, Year 2003, for in part supporting this research.

References

[1] R. L. Hsu, M. Abdel-Mottaleb and A. K. Jain, "Face Detection in Color Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.2, No.5, pp.696-706, May 2002.

[2] J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink and G. Sagerer, "Improving Adaptive Skin Color Segmentation by Incorporating Results from Face Detection," *IEEE Int. Workshop on Robot and Human Interactive Communication*, September 2002.

[3] N. Tanibata, N. Shimada and Y. Shirai, "Extraction of Hand Features for Recognition of Sign Language Words," *Proc. of Int. Conf on Vision Interface,* pp.391-398, 2002.

[4] K. Imagawa, S. Lu and S. Igi, "Color-based Hands Tracking System for Sign Language Recognition", *Face and Gesture (FG 1998)*, pp.462-467,1998.

[5] K. Jack, *Video Demystified A Handbook for the Digital Engineer, 3rd edition*, LLH Technology Publishing, 2001.

[6] J. Yang and A. Waibel, "A Real-Time Face Tracker," *Proc. of Third Workshop on Applications of Computer Vision*, pp.142-147,1996.

[7] C. R. Wren, A. Azarbayejani, T. Darrel, and A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp.780-785, July 1997.

[8] M. C. Shin, K. I. Chang and L. V. Tsap "Does Colorspace Transformation Make Any Difference on Skin Color," *IEEE Workshop on Applications of Computer Vision*, 2002.

[9] Thai sign language video sequences, Ministry of Education.

การตรวจหาและติดตามใบหน้าและมือสำหรับวีดิทัศน์ภาษามือ

Face and Hands Detection and Tracking for Sign Language Video

นรุตม์ สุนทรานนท์[1] สุภาวดี อร่ามวิทย์[1] และ ธนารัตน์ ชลิดาพงศ์[2]

[1]ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย
254 ถนนพญาไทย เขตปทุมวัน กรุงเทพฯ 10330
โทร. 0-2218-6909 E-mail: Supavadee.A@chula.ac.th

[2]คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
3 หมู่ 2 ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพฯ 10520
โทร. 0-2737-2551 ต่อ 526 E-mail: thanarat@it.kmitl.ac.th

บทคัดย่อ

บทความฉบับนี้นำเสนอ ระบบตรวจหาและติดตามใบหน้าและมือเพื่อประโยชน์ในการแปลความหมายภาษามือไทย ในการเลือกปริภูมิสีที่เหมาะสม สำหรับการแยกสีผิวมนุษย์ ได้ใช้วิธีการวัดสัญญาณรบกวนด้วยแบบจำลองชนิดไม่มีตัวแปรของปริภูมิสีต่าง ๆ และเลือกที่จะใช้ปริภูมิสี CbCr ในการทำงานจริงจะใช้แบบจำลองชนิดวงรี เพื่อลดความซับซ้อนของการคำนวณ เมื่อได้บริเวณสีผิวกระบวนการถัดมาก็จะทำการหาคุณลักษณะสำคัญต่าง ๆ เช่น บริเวณใบหน้า หรือ การแยกระหว่างมือกับแขน ในส่วนการติดตาม ใช้วิธีการจับคู่บล็อกที่มีค่าผิดพลาดกำลังสองเฉลี่ยต่ำสุดเพื่อหาตำแหน่งในเฟรมถัดไป และได้แสดงถึงความสามารถของระบบในการตรวจหาส่วนสำคัญและติดตามวีดิทัศน์ภาษามือในส่วนผลการทดลอง

Abstract

In this paper, we develop the face and hand detection and tracking for sign language recognition system. We first perform preliminary evaluation on several color metrices to find the most suitable one using non-parametric model approach. Then, we propose to use the elliptical model on CbCr to lower the complexity of the detection algorithm and to better model the skin color. After the skin regions from the input video have been segmented, the interested facial features and hands are detected using luminance differences and skeleton features respectively. In the tracking stage, each blob determines search region and find MMSE (Minimum Mean Square Error) to match its own blob. The block matching method between previous and current frame is used. Experimental results show that our proposed system is able to detect and track face and hands of sign language video sequences.

Keywords: Sign language, Skin-color, Face detection, Tracking.

1. คำนำ

เนื่องจากปัญหาในการสื่อสารระหว่าง ผู้ที่มีปัญหาด้านการได้ยินกับคนปกติ มีอุปสรรคบ่อยครั้ง ดังนั้น จึงมีความต้องการระบบที่ใช้เป็นตัวกลางในการแปลความหมายภาษามือ เพื่อให้เกิดความเข้าใจที่ตรงกันแก่ทั้งสองฝ่าย ทั้งยังเป็นเครื่องมือที่ช่วยให้ผู้ที่มีปัญหาด้านการได้ยินมีโอกาสใช้ติดต่อสื่อสารกับคนปกติในสังคมภายนอกมากยิ่งขึ้น อันจะเป็นประโยชน์ต่อการสนองตอบความต้องการ หรือ นำแนวความคิดของพวกเขาเหล่านั้นมาร่วมกันพัฒนาประเทศชาติต่อไป

การสื่อสารโดยใช้ภาษามือของแต่ละประเทศ ก็มีลักษณะแตกต่างกันเช่นเดียวกับภาษาพูด ด้วยเหตุผลนี้ระบบการแปลภาษามือที่นำเสนอในต่างประเทศ [3,4] จึงไม่สามารถนำมาใช้ในการแปลภาษามือไทยได้ ฉะนั้น เรามีความจำเป็นต้องสร้างระบบนี้ขึ้นเอง และจากการศึกษาเบื้องต้นเกี่ยวกับภาษามือไทย ผู้ที่มีปัญหาด้านการได้ยินนอกจากจะรับรู้ข่าวสารหรือข้อมูลของคู่สนทนาโดยการแสดงท่าทางของมือแล้ว ยังต้องอาศัยการสังเกตอากัปกิริยาของคู่สนทนาผ่านทางใบหน้าอีกด้วย

บทความนี้จึงนำเสนอ ระบบตรวจหาและติดตามใบหน้าและมือเพื่อนำไปใช้ประโยชน์ในการแปลความหมายภาษามือไทย โดยระบบประกอบไปด้วย 2 ส่วน คือ ส่วนเริ่มต้น ซึ่งทำการตรวจหาตำแหน่งส่วนต่าง ๆ ที่สำคัญของผู้แปลออกมา และ ส่วนติดตาม จะใช้ตำแหน่งที่ได้มาเพื่อค้นหาแบบจับคู่แบบบล็อกในเฟรมถัดไป

2. การแยกบริเวณสีผิวมนุษย์

ในการพิจารณาถึงวิธีตรวจหาบริเวณใบหน้าและมือ เราจะนำคุณลักษณะเด่นที่สำคัญอันหนึ่งมาใช้ คือ สีผิวมนุษย์ แต่ปัญหาที่พบคือค่าจากสัญญาณวีดิทัศน์เป็นค่า RGB ไม่เหมาะสมในการแยกสีผิว เพราะมีการรวมกันขององค์ประกอบความสว่างและสี เมื่อมีการเคลื่อนที่ในส่วนใบหน้าและมือของผู้แปลขณะกำลังแสดงท่าทาง แสงที่ตกกระทบจะมีค่าเปลี่ยนแปลงอยู่ตลอดเวลา ทำให้แบบจำลองตีความหมายผิดพลาด ดังนั้นเราจึงต้องทำการแปลงไปสู่ปริภูมิสีที่เหมาะสมกว่า กล่าวคือมีการแยกกันระหว่างความสว่างกับสี และเหมาะสมต่อการแยกสีผิว

2.1 ปริภูมิสี

ปัจจุบันมีผู้เสนอวิธีการแยกสีผิวมนุษย์ในปริภูมิสีต่าง ๆ ไว้อย่างมากมาย เพื่อให้เกิดความชัดเจนถึงความถูกต้องของการแยกสีผิวในแต่ละปริภูมิสี บทความนี้จึงเสนอวิธีการใช้แบบจำลองชนิดไม่มีตัวแปร (Non-parametric Model) วัดผลกระทบของสัญญาณรบกวนเปรียบเทียบใน 3 ปริภูมิสี ที่มักถูกนำมาใช้บ่อย คือ CbCr [1,5], rg [2] และ HS [3]

- ปริภูมิสี YCbCr

เป็นปริภูมิสีดิจิทัลที่กำหนดขึ้นโดย BT.601 [6] ซึ่งถูกนำมาใช้ในมาตรฐานของการบีบอัดข้อมูลวีดีโอ เช่น H.26L, H.264, MPEG-4 ฯลฯ

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.2568 & 0.5041 & 0.0980 \\ -0.1482 & -0.2910 & 0.4392 \\ 0.4392 & -0.3678 & -0.0714 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \quad (1)$$

การแปลงจากปริภูมิสี RGB ไปสู่ YCbCr ทำโดยใช้สมการที่ 1 ซึ่งค่า Y คือ ค่าองค์ประกอบความสว่าง มีค่าตั้งแต่ 16 – 235 ส่วนค่า Cb และ Cr คือ ค่าองค์ประกอบของสี จะแสดงถึงความแตกต่างระหว่าง สีน้ำเงินกับความสว่าง และ สีแดงกับความสว่าง ตามลำดับ มีค่าตั้งแต่ 16 – 240

- ปริภูมิสี rg

ถือเป็นปริภูมิสีหนึ่งซึ่งให้ค่าองค์ประกอบของสีอย่างเดียว โดยค่า r และ g คือ ค่าที่ได้จากการนอร์แมลไลซ์ มีค่าอยู่ระหว่าง 0 – 1

$$r = \frac{R}{R+G+B} \quad (2)$$

$$g = \frac{G}{R+G+B} \quad (3)$$

- ปริภูมิสี HSV

ปริภูมิสี HSV ค่าของ H และ S คือ ค่าองค์ประกอบของสี และค่า V คือ ค่าองค์ประกอบความสว่าง โดยที่ H มีค่าตั้งแต่ 0 – 360 องศา ส่วนค่า S และ V มีค่าตั้งแต่ 0 – 1 การแปลงจาก RGB ไปสู่ HSV ทำได้ดังนี้

$$H = \begin{cases} H_1 & ; B \le G \\ 360^\circ - H_1 & ; B > G \end{cases} \quad (4)$$

โดยที่

$$H_1 = \cos^{-1}\left\{ \frac{0.5[(R-G)+(R-B)]}{\sqrt{(R-G)(R-G)+(R-B)(G-B)}} \right\} \quad (5)$$

$$S = \frac{\max(R,G,B) - \min(R,G,B)}{\max(R,G,B)} \quad (6)$$

$$V = \frac{\max(R,G,B)}{255} \quad (7)$$

2.2 การแยกบริเวณสีผิวโดยใช้แบบจำลองชนิดวงรี

ในส่วนนี้จะอธิบายถึง ปริภูมิสีและแบบจำลองในการแยกสีผิวที่เลือกนำมาใช้ในระบบ โดยต้องคำนึงถึงความซับซ้อนของการคำนวณที่ต่ำ เพื่อให้สามารถทำงานได้ในเวลาจริง

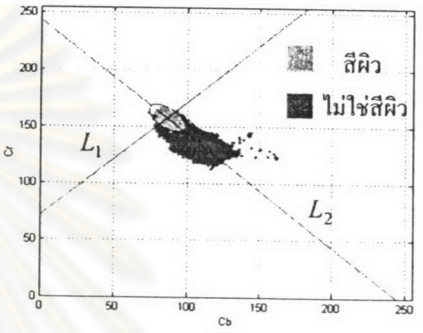
เมื่อเปรียบเทียบผลของสัญญาณรบกวนโดยใช้แบบจำลองชนิดไม่มีตัวแปร (ตารางที่ 1) อีกทั้งยังมีข้อได้เปรียบในเรื่องการนำไปผนวกใช้งานร่วมกับมาตรฐานการบีบอัดวีดิทัศน์ จึงเลือกการแปลงค่าไปสู่ปริภูมิสี YCbCr สำหรับวิธีที่จะแก้ปัญหาเรื่องความไม่แน่นอนของแสงที่ตกกระทบบนใบหน้าและมือมีค่าเปลี่ยนแปลงอยู่ตลอดเวลา แบบจำลองที่ใช้แยกส่วนสีผิวจะพิจารณาเพียงองค์ประกอบสี (CbCr) โดยละทิ้งองค์ประกอบทางความสว่าง (Y) ไป และเมื่อสังเกตการกระจายตัวของค่าบริเวณสีผิว (รูปที่ 1) จะเกาะกลุ่มกันเป็นลักษณะวงรี [1] เป้าหมายคือเราจะทำการแยกค่าส่วนนี้ออกมา ซึ่งเราสามารถสร้างแบบจำลองวงรีได้ดังนี้

$$L_1 : AC_b + BC_r + C = 0 \quad (8)$$

$$L_2 : BC_b - AC_r + D = 0 \quad (9)$$

โดยที่ L_1 และ L_2 คือ แกนเอกและแกนโทของวงรี ตามลำดับ

A, B, C และ D สามารถหาได้จากค่าเฉลี่ย และ ค่าความแปรปรวนร่วมของ CbCr



รูปที่ 1 การกระจายตัวของจุดภาพบริเวณสีผิวในปริภูมิสี CbCr

$$\frac{L_1^{\,2}}{a^2} + \frac{L_2^{\,2}}{b^2} = 1 \quad (10)$$

โดยที่ a และ b คือ ความยาวของแกนเอก และแกนโทของวงรี ตามลำดับ

ระบบจะกำหนดจุดภาพบริเวณสีผิวโดยตัดสินจากค่า CbCr ที่ถูกล้อมรอบอยู่ภายในวงรี

$$MSkin(i,j) = \begin{cases} 1, & if\ enclosed\ with\ ellipse \\ 0, & otherwise \end{cases} \quad (11)$$

โดยที่ $MSkin(i,j)$ คือ มาสก์ของบริเวณสีผิว

i และ j คือ ความสูงและความกว้างของภาพ มีค่าสูงสุดเท่ากับ 240 และ 320 ตามลำดับ

3. การตรวจจับบริเวณสำคัญเพื่อประโยชน์ในการแปล

3.1 ส่วนสำคัญบนใบหน้า

- การแยกบริเวณใบหน้าและคอ

จากขั้นตอนการแยกบริเวณส่วนสีผิว เมื่อได้ตำแหน่งของกลุ่มสีผิวส่วนใบหน้าแล้ว ในหัวข้อนี้ จะแสดงถึงวิธีการในการแยกระหว่างบริเวณของใบหน้าและคอ (รูปที่ 2) ออกจากกัน ดังนี้

- การตรวจหาขอบด้วยวิธีการแบบไขว้ศูนย์ (zero-cross)
- เติมค่าสีขาวลงในบริเวณที่ถูกล้อมรอบเป็นวงปิด
- การวิเคราะห์หาบริเวณเชื่อมต่อกัน และทำการเลือกบริเวณที่ใหญ่ที่สุด เพื่อกำหนดบริเวณนั้นเป็นใบหน้า

(ก) (ข) (ค)

รูปที่ 2 (ก) การตรวจหาขอบด้วยวิธีแบบไขว้ศูนย์
(ข) เติมค่าสีขาวลงในบริเวณที่ถูกล้อมรอบ
(ค) ขอบเขตบริเวณใบหน้า

- การแยกส่วนตา ปาก และจมูก

เมื่อพิจารณาบริเวณใบหน้า ค่าจุดภาพของส่วนตา ปาก และจมูก จะประกอบไปด้วยจุดภาพที่มี ความเข้มของแสงน้อยกว่าส่วนอื่น ๆ บนใบหน้า (ค่า Y ต่ำ) จึงแยกบริเวณเหล่านี้โดยใช้ค่าตัดสิน ตามสมการที่ 12

$$TH1 = Mean(Y(i_f, j_f)) - \beta * Std(Y(i_f, j_f)) \quad (12)$$

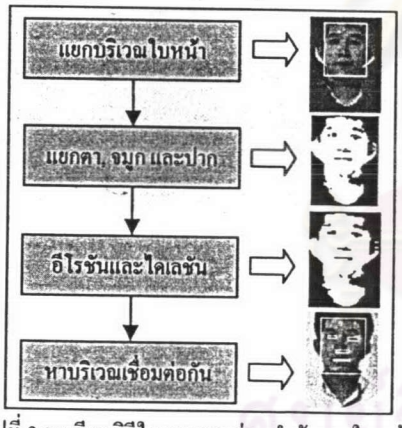
ซึ่ง $Y(i_f, j_f)$ คือ ค่าองค์ประกอบความสว่างบริเวณสีผิว

β คือ ค่าถ่วงน้ำหนัก ในที่นี้มีค่าเท่ากับ 1.5

$$MFacial(i_f, j_f) = \begin{cases} 0, if\ \ MSkin(i_f, j_f) \le TH1 \\ 1, if\ \ MSkin(i_f, j_f) > TH1 \end{cases} \quad (13)$$

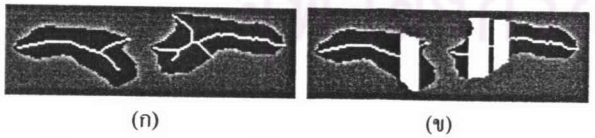
โดยที่ $MFacial(i_f, j_f)$ คือ มาสก์ของส่วนสำคัญบนใบหน้า

i_f และ j_f คือ กลุ่มของจุดภาพบริเวณสีผิว



รูปที่ 3 ระเบียบวิธีในการแยกส่วนสำคัญบนใบหน้า

3.2 การแยกบริเวณส่วนแขนและมือ



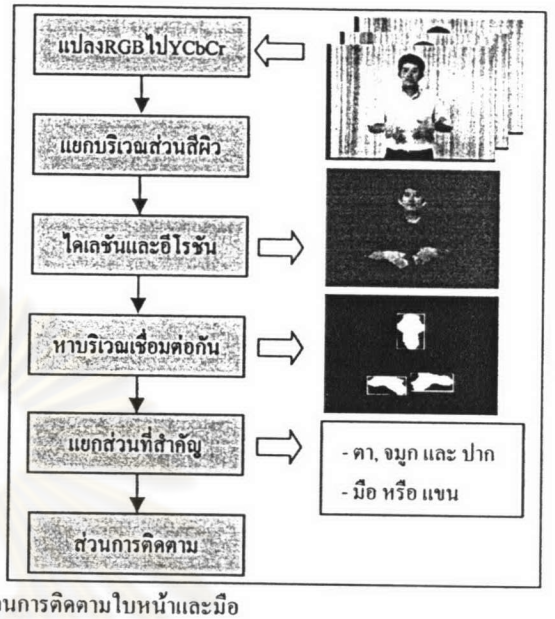
(ก) (ข)

รูปที่ 4 (ก) มาสก์บริเวณมือที่ผ่านกระบวนการธินนิ่ง (Thinning)
(ข) กำหนดขอบเขตของมือ

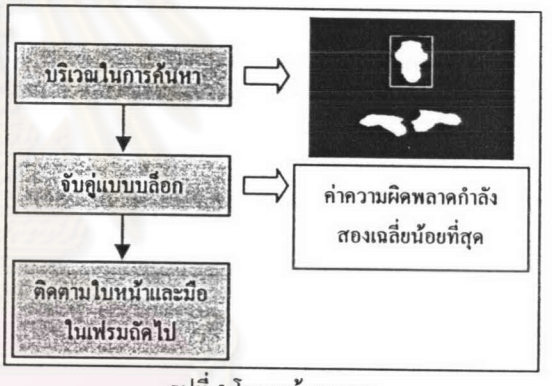
กรณีที่ผู้แปลภาษามือสวมเสื้อแขนสั้น จำเป็นที่จะต้องมีการแยกระหว่างส่วนมือกับแขนออกจากกัน ซึ่งวิธีการที่นำมาใช้ คือ การทำธินนิ่งส่วนการตัดสินว่าที่เป็นมือหรือแขน จะถือเอาส่วนที่มีรอยแยกมากกว่าหลังจากทำธินนิ่งเป็นมือ

4. โครงสร้างระบบ

- ส่วนเริ่มต้น



- ส่วนการติดตามใบหน้าและมือ



รูปที่ 5 โครงสร้างระบบ

ระบบที่นำเสนอแบ่งเป็น 2 ส่วนด้วยกัน คือ ส่วนเริ่มต้น และส่วนการติดตามใบหน้าและมือ โดยส่วนเริ่มต้นจะแปลงสัญญาณวีดิทัศน์จาก RGB ไปสู่ YCbCr จากนั้นจึงทำการแยกบริเวณส่วนสีผิว และผ่านตัวดำเนินการแบบปิด (ไดเลชันและอีโรชัน) การวิเคราะห์หาบริเวณเชื่อมต่อกัน กระบวนการในการแยกส่วนสำคัญต่าง ๆ ที่เป็นประโยชน์ในการนำไปใช้แปลความหมายภาษามือ หลังจากนั้นจะเป็นส่วนของการติดตาม

ระเบียบวิธีในการติดตาม ในเฟรมแรกระบบจะทำการตรวจหาองค์ประกอบต่าง ๆ ที่จำเป็น หลังจากนั้นเฟรมถัดมาระเบียบการติดตามส่วนต่าง ๆ เหล่านั้น จะถูกนำมาใช้ ซึ่งวิธีที่ใช้ คือการจับคู่บล็อกที่มีค่าผิดพลาดเฉลี่ยกำลังสองน้อยที่สุด

$$(X, Y) = \arg \min_{dx, dy \in S} Sum(DFD) \quad (14)$$

$$DFD = \frac{1}{N} \{I_{t-1}(x, y) - I_t(x + dx, y + dy)\}^2 \quad (15)$$

ซึ่ง DFD คือ ความแตกต่างของระยะกระจัดระหว่างเฟรม,

I_{t-1} คือ เฟรมก่อนหน้า, I_t คือ เฟรมปัจจุบัน, N คือ ขนาดของบล็อก , dx และ dy คือ ขอบเขตในการค้นหา

5. ผลการทดลอง

ประสิทธิภาพของการแยกส่วนสีผิว จะแสดงโดยใช้พารามิเตอร์ 2 ตัว คือ อัตราการตรวจหา (Detection rate: DR) และ อัตราการเตือนความผิดพลาด (False alarm rate: FAR) ค่า DR ที่สูง และ FAR ต่ำ จะให้ผลที่ดี

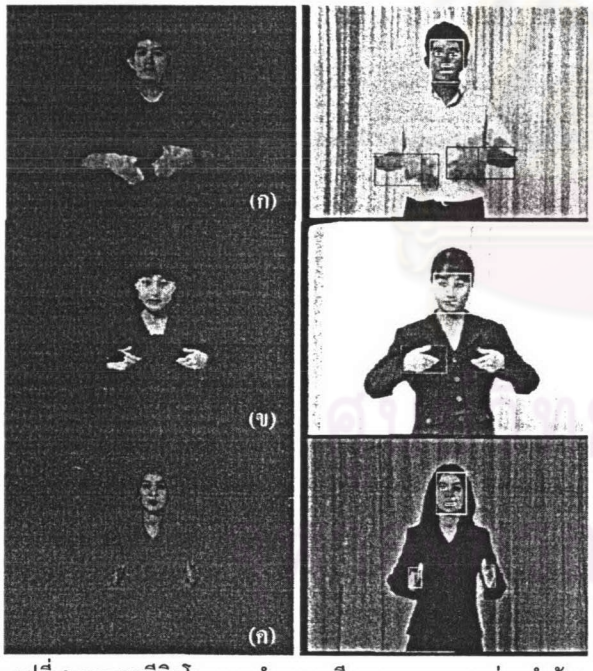
$$DR = \frac{TP}{TP + FN} \times 100\% \quad (16)$$

$$FAR = \frac{FP}{TP + FP} \times 100\% \quad (17)$$

โดยที่ TP ย่อจาก True Positive, FP ย่อจาก False Positive และ FN ย่อจาก False Negative

ตารางที่ 1 เปรียบเทียบประสิทธิภาพในแต่ละปริภูมิสี ด้วยแบบจำลองชนิดไม่มีตัวแปร ในรูปภาพทดสอบ ก, ข และ ค [7] ตามลำดับ

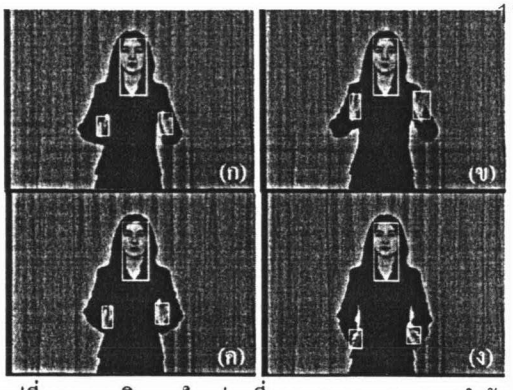
ปริภูมิสี	DR(ก)	FAR(ก)	DR(ข)	FAR(ข)	DR(ค)	FAR(ค)
CbCr	68.85	1.67	59.62	2.65	68.27	0.95
rg	64.55	6.45	64.72	70.54	59.45	81.86
HS	52.89	8.54	48.75	68.74	55.45	83.24



รูปที่ 6 การแยกสีผิวโดยแบบจำลองวงรีและการตรวจหาส่วนสำคัญ

ตารางที่ 2 ประสิทธิภาพแต่ละรูปเมื่อแยกสีผิวด้วยแบบจำลองชนิดวงรี

รูปภาพ	DR	FAR
ก	91.15	4.27
ข	79.60	4.15
ค	96.26	7.94



รูปที่ 7 ผลการติดตามในเฟรมที่ 10,20,40 และ 70 ตามลำดับ

6. สรุป

บทความนี้เสนอ ระบบการตรวจหาและติดตามใบหน้าและมือ เพื่อนำไปใช้ในระบบแปลความหมายภาษามือไทย โดยขั้นต้นเราทำการเปรียบเทียบเพื่อหาปริภูมิสีที่เหมาะแก่การแยกสีผิว ด้วยแบบจำลองชนิดไม่มีตัวแปร พบว่าปริภูมิสี CbCr ให้ผลที่ดีกว่าอีก 2 ปริภูมิสี (ตารางที่ 1) สำหรับการใช้งานจริงในระบบ เพื่อให้มีการคำนวณไม่ซับซ้อน จึงเลือกแบบจำลองชนิดวงรีบนปริภูมิสี CbCr จากนั้นจะแยกส่วนต่าง ๆ ที่สำคัญบนใบหน้าโดยใช้ความแตกต่างทางความสว่าง และกระบวนการธินิ่งในการแยกระหว่างมือกับแขน เพื่อทำการติดตามแต่ละส่วนต่อไป

7. กิตติกรรมประกาศ

ขอขอบคุณโครงการวิจัยร่วมภาครัฐและเอกชนปี 2546 ภาควิชาวิศวกรรมไฟฟ้า จุฬาลงกรณ์มหาวิทยาลัย ที่ให้การสนับสนุนงานวิจัยนี้

เอกสารอ้างอิง

[1] R. L. Hsu, M. Abdel-Mottaleb and A. K. Jain, "Face Detection in Color Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.2, No.5, pp.696-706, May 2002.

[2] J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink and G. Sagerer, "Improving Adaptive Skin Color Segmentation by Incorporating Results from Face Detection," *IEEE Int. Workshop on Robot and Human Interactive Communication*, September 2002.

[3] N. Tanibata, N. Shimada and Y. Shirai, "Extraction of Hand Features for Recognition of Sign Language Words," *Proc. of Int. Conf on Vision Interface*, pp.391-398, 2002.

[4] K. Imagawa, S. Lu and S. Igi "Color-based Hands Tracking System for Sign Language Recognition," *FG 1998*, pp.462-467,1998.

[5] J. Yang and A. Waibel, "A Real-Time Face Tracker," *Proc. of Third Workshop on Applications of Computer Vision*, pp.142-147,1996.

[6] K. Jack, *Video Demystified A Handbook for the Digital Engineer, 3rd edition, LLH Technology Publishing*, 2001.

[7] วีดิทัศน์ภาษามือไทย สมาคมคนหูหนวกแห่งประเทศไทย

Improved Face and Hand Tracking for Sign Language Recognition

N. Soontranon and S. Aramvith*, T.H. Chalidabhongse†

*Department of Electrical Engineering
Chulalongkorn University
Bangkok 10330 Thailand
Tel: +66-2218-6909
E-mail: Supavadee.A@chula.ac.th

†Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520 Thailand
Tel: +66-2737-2551 Ext.526
E-mail: thanarat@it.kmitl.ac.th

Abstract

In this paper, we develop face and hand tracking for sign language recognition system. The system is divided into two stages; the initial and tracking stages. In initial stage, we use the skin feature to localize face and hands of signer. The ellipse model on CbCr space is constructed and used to detect skin color. After the skin regions have been segmented, face and hand blobs are defined by using size and facial feature with the assumption that the movement of face is less than that of hands in this signing scenario. In tracking stage, the motion estimation is applied only hand blobs, in which first and second derivative are used to compute the position of prediction of hands. We observed that there are errors in the value of tracking position between two consecutive frames in which velocity has changed abruptly. To improve the tracking performance, our proposed algorithm compensates the error of tracking position by using adaptive search area to re-compute the hand blobs. Simulation results indicate our proposed algorithm can track face and hand with greater precision with negligible computational complexity increase.

Keywords: *Hand tracking, sign language, ellipse model, skin-color segmentation, motion estimation.*

1. Introduction

Information and knowledge are expanding in quantity and accessibility. However, people with functional limitations, such as deaf people, often experience wide communication gaps. Due to the hearing limitation, deaf people have developed their own culture and methods for communicating among them as well as with hearing groups by rely on signing. The sign language translator is thus an important tool to enabling effective communication between deaf and hearing people. There are several researches in sign language recognition such as Japanese SLR [1,3], Korean SLR [4] and etc. Nowaday, there are two ways to collect gesture data for recognition. First, device-based measurement technique measures hand gestures with equipment such as data gloves which can archive the accurate position of hand gestures as its position is directly measured. Nevertheless, the gloves are very expensive and it cannot be used to collect facial gesture data. Second is the vision-based approach. This technique can cover both face and hands signer in which signer does not need to wear data gloves device. All processing tasks are solved by using computer vision techniques which are more flexible and useful than prior approach. In this way, the skin color regions are segmented to locate the position of finger, hand and face. This vision-based technique can track face and hands during signing to recognize the movement. In addition hand tracking and gesture recognition have been widely addressed. Kalman filter-based dynamic model is used to track the movement of the hands using the temporal synchronization by the velocity and acceleration of each hand [2].

In this work, we develop the face and hand detection and tracking for sign language recognition system. The system is divided into 2 stages; the initial and tracking stages. The initial stage has been discussed extensively in [5]. In initial stage, skin color feature is selected. After the skin regions have been segmented, the interested to separate face or hands using size and facial feature. Because the movement of face is less than that of hands, we propose to map the motion estimation model only hand blobs in the tracking stage. For face tracking, it is easier to define the search region based on the position in the previous frame. To improve the tracking precision, the tracking method can compensate error of using adaptive search area to re-compute the hands blobs iteratively. Simulation

results indicate our proposed algorithm can track face and hand with greater precision with negligible computational complexity increase.

The organization of this paper is as follows. Section 2, we describe the initial stage consisted of detection and localization face and hand. Section 3 presents the tracking stage of face and hand. Section 4 shows the experimental results. Section 5 concludes the paper.

2. Initial Stage

The purposes of initial stage are to locate and to track the signer's face and hand by segmentation of the regions from the sign language sequence using human skin color model.

2.1. Skin color Segmentation

In this section, we describe the classification method employed to classify pixels as skin or non-skin. The skin color segmentation confuses under different lighting conditions while signer is moving his face and hand, thus it is important to choose an appropriate color space for skin segmentation. Popular color spaces have been proposed are the RGB [9], normalized RGB [8], HSV [1] and YCbCr [5,7]. Considering the result of non-parametric skin color modelling [5] which compares the performance of skin color detection, the experimental result suggested us to use the YCbCr color space.

To construct the skin color model, we manually crop training skin regions, plot them on CbCr plane (use only chrominance component that neglect luminance component), as shown in Fig. 1(a), and segment face and hands using ellipse skin color model [5]. Ellipse model can be derived from eccentricity method by equation (1)-(4). The resulting skin color mask is shown in Fig. 1(c).

$$\begin{bmatrix} \alpha(Cb,Cr) \\ \beta(Cb,Cr) \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} Cb \\ Cr \end{bmatrix} \quad (1)$$

,where α, β are the transformation domain and θ can be obtained from eqs. (2) and (3).

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20}+\mu_{02}}\right) \quad (2)$$

$$\mu_{pq} = \sum_{i=1}^{n}\sum_{j=1}^{n}(Cb_i - \overline{Cb})^p (Cr_j - \overline{Cr})^q \quad (3)$$

$$\frac{L_1^{\,2}}{a^2} + \frac{L_2^{\,2}}{b^2} = 1 \quad (4)$$

,where a is the size of ellipse at major axis, b is the size of ellipse at minor axis, L_1 and L_2 are the major and minor axis of the ellipse respectively.

Note that a, b can be obtained from mean and standard deviation of (α, β), in which L_1 and L_2 intersect and perpendicular at mean (C_b, C_r). The tangent of θ is used to slope of the L_1. The optimized threshold is derived from ground truth.

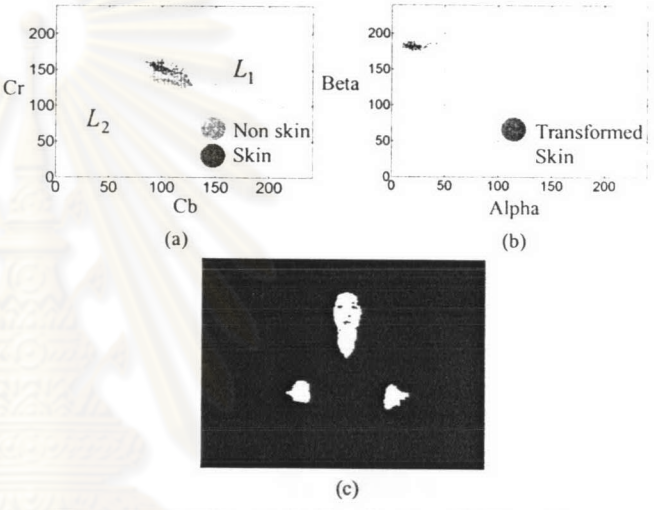


Figure 1. (a) Skin pixels distribution. (b) Transform skin pixels to α, β. (c) Skin color mask.

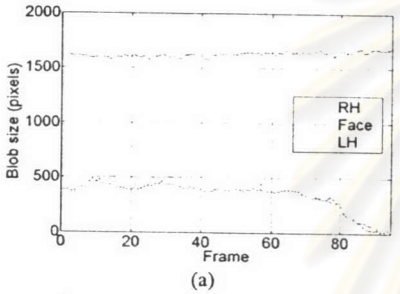
2.2. Face and hand detection

After the human skin color has been segmented, the noises are eliminated by morphological operator. Feature detection is applied next to separate between face and hands blobs by using size and facial feature. We need to detect the face and hand because, in order to recognize the sign language, the position of face and hands must be analyzed and interpreted. Thus, in the tracking stage, the located face and hand blobs will be mapped using different motion estimation model.

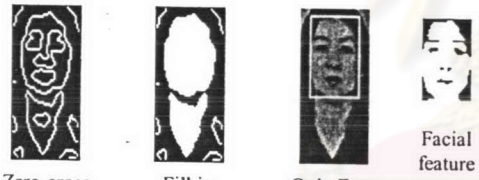
Fig. 2(a) shows the difference between size in pixels of face and hand blobs. Connected component of face including the neck region contains 1,600 pixels but hand blobs contain 400 pixels. Note that, for the frame number 80-95, the size of hand blobs is close to zero because both hands are moving out of frame. From the curve, it is evident that there are much different

between blob size of face and hands. Thus, the threshold could be selected correspondingly.

However, in the case where the magnitude of blob size of face and hands are close, for example, in the case of signer wear short sleeve in which the connected component of hand may be bigger than face, the facial feature process needs to be added. Fig. 2(b) shows the facial feature process. First, face region is separated from the head area using zero-cross edge detection, then fills the holes in enclosed boundary. The face region is defined as the biggest component after performing the connected component analysis. After that, facial features such as eyes, nose, and mouth are detected based on the fact that these areas contain stronger luminance component than other face regions.



(a)



(b)

Figure 2. (a) Blobs size (b) Facial features.

Block diagram of initial stage is shown in Fig. 3. The input RGB sequence is transformed to YCbCr color space. The skin regions are first segmented using skin color feature by elliptical model on CbCr. The segmented skin-color regions are closed, i.e., dilation and erosion, using digital disc size 5 to fill up the small holes. Next, the connected component analysis is applied to eliminate some noise regions and to localize the salient skin color blobs. Then, the features detection explained in next section is performed to initially specify the positions of those interested features needed in further tracking.

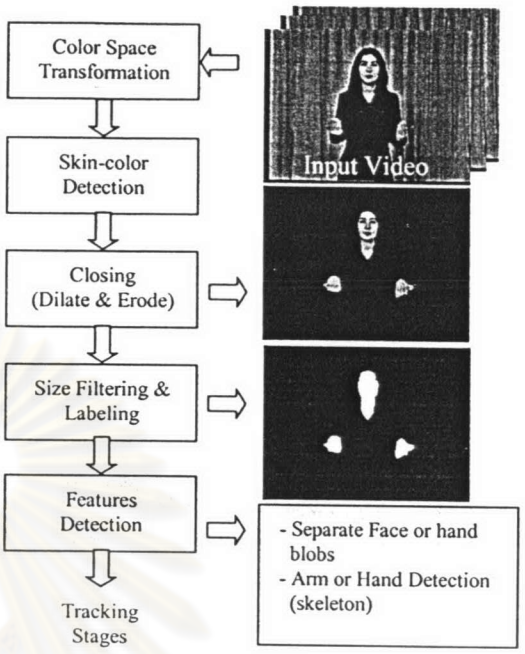
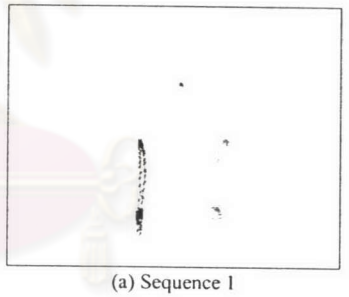
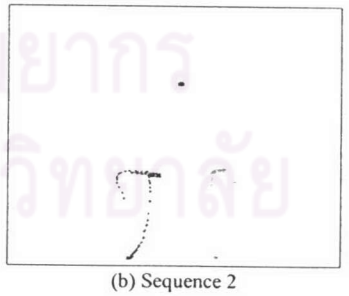


Figure 3. Initial stage.

3. Tracking Stage



(a) Sequence 1



(b) Sequence 2

Figure 4. Trajectory of each blobs.

After obtaining the localization of face and hand from the initial stage, in the tracking stage, the system predicts the new location of face and hand in the next frame by mapping the estimation model for them. This

is done to lower the complexity of the system. In the tracking stages, the new motion estimation model has been proposed. Two sequences are used in the simulations, as their trajectories are shown in Fig. 4. In both sequence, we can see that the face blob almost stands still while the trajectory of the hands in the first sequence moves vertically and the trajectory of the hands in the second sequence moves horizontally.

3.1. Face Tracking

Since the face blob hardly changes its position, it is easily to determine face region using the position from previous frame. The idea is to expand the search region from the former position, shown in eq. (5).

$$\begin{bmatrix} x_{Search}(i+1) \\ y_{Search}(i+1) \end{bmatrix} = \begin{bmatrix} x(i) \pm W_i \\ y(i) \pm W_i \end{bmatrix} \quad (5)$$

,where W_i is expanded search window in the next frame. Note that the value of W_i = 10 is used in the experiment.

In the following section, we discuss the hand tracking which use the location in the previous frames to derive the motion estimation model for predicting the future position.

3.2. Hand Tracking

For hand tracking, motion estimation is constructed by using velocity and acceleration among consecutive frames. The previous location of hands is used to predict the future location of blobs. The location of every pixels of each blob is calculated from the centroid of each blob. In the first three frames, the locations of the hand blobs are kept such that we can calculate both velocity and acceleration.

Velocity and acceleration of horizontal and vertical between frame i-1 and i can be computed from eqs. (6)-(7).

$$\begin{bmatrix} V_x(i) \\ V_y(i) \end{bmatrix} = \begin{bmatrix} x(i) \\ y(i) \end{bmatrix} - \begin{bmatrix} x(i-1) \\ y(i-1) \end{bmatrix} \quad (6)$$

$$\begin{bmatrix} A_x(i) \\ A_y(i) \end{bmatrix} = \begin{bmatrix} V_x(i) \\ V_y(i) \end{bmatrix} - \begin{bmatrix} V_x(i-1) \\ V_y(i-1) \end{bmatrix} \quad (7)$$

Motion estimation can be computed as in eq. (8), also shown in Fig. 5.

$$\begin{bmatrix} x(i+1) \\ y(i+1) \end{bmatrix} = \begin{bmatrix} x(i)+V_x(i) \\ y(i)+V_y(i) \end{bmatrix} + \frac{1}{2}\begin{bmatrix} A_x(i) \\ A_y(i) \end{bmatrix} \quad (8)$$

In this paper, we proposed the motion estimation model as shown in eq. (9) using additional information from the velocity of previous frame. It will be shown later that this model helps reduce the prediction error of hand position.

$$\begin{bmatrix} x(i+1) \\ y(i+1) \end{bmatrix} = \begin{bmatrix} x(i) \\ y(i) \end{bmatrix} + \frac{1}{2}\begin{bmatrix} V_x(i)+V_x(i-1)+A_x(i) \\ V_y(i)+V_y(i-1)+A_y(i) \end{bmatrix} \quad (9)$$

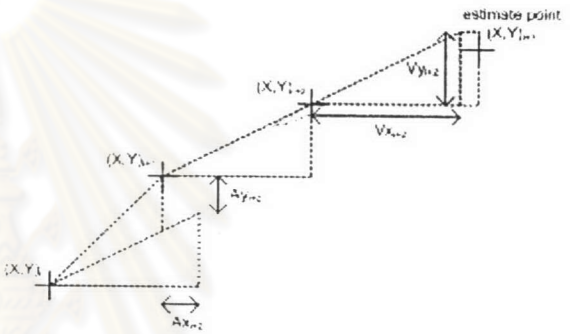


Figure 5. The method of motion estimation.

3.3. Adaptive Search Window based on Velocity

As we observed that there are errors in the value of tracking position between two consecutive frames in which velocity has changed abruptly, our proposed algorithm compensates the error of tracking position by using adaptive search area to re-compute the hand blobs. The search window will be expanded in the case where acceleration exceeds a threshold, as shown in eq. (10).

$$W_i = \begin{cases} W_{small} & \text{if } A_{x,y} < \varepsilon \\ W_{large} & \text{if } A_{x,y} \geq \varepsilon \end{cases} \quad (10)$$

W_{small} : small window for search region.

W_{large} : large window for search region.

ε : threshold value

Fig. 7 summarizes the process of tracking stage.

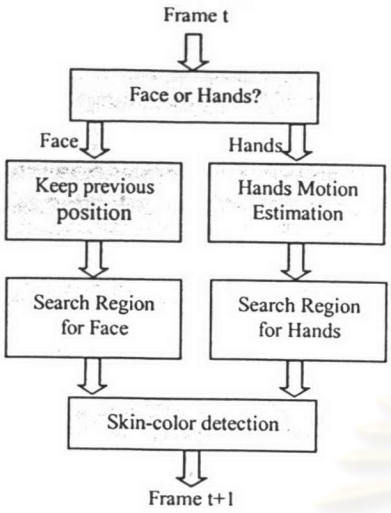


Figure 7. Tracking stage.

4. Experimental Results

In this section, we demonstrate the results of our proposed algorithm on Thai Sign Language sequence (.avi format) [11] which has frame rate 15 fps and format of 320x240 pixels.

Fig. 8 shows the result of prediction error of right hand blobs. Fig. 8(a) shows the velocity of each frame. The circle highlights the change of the hand position abruptly. This leads to the prediction error, as can be seen in Fig. 8(b). By using our proposed method, the prediction error is minimized, as shown in Fig. 8(c). Table 1 summarizes the improvement of prediction error. Table 2 presents prediction error and processing time between normal tracking, (eq.8), versus proposed adaptive window mode. The experimental results indicate that our proposed method is able to decrease the prediction error up to 96.87% with negligible increase in computational complexity of up to 4%.

Table 1 Comparative performance of motion estimation model in two Thai sign language sequences.

Video Sequence	*Method 1 Eq. 8*	*Method 2 (proposed)*
	Position error (pixels)	Position error (pixels)
Seq.1	7.06 (RH) 9.97 (LH)	0.80 (RH) 1.80 (LH)
Seq.2	18.37 (RH) 11.50 (LH)	2.47 (RH) 0.36 (LH)

Note RH: Right Hand, LH: Left Hand

Table 2 Position error and processing time between normal tracking, eq. (8), versus proposed adaptive window mode.

Video	*Decrease Position Error (%)*	*Increase Processing Time (%)*	*Number of iterative computation*
Seq.1	88.67 (RH) 81.95 (LH)	2.45	8
Seq.2	86.55 (RH) 96.87 (LH)	4.05	16

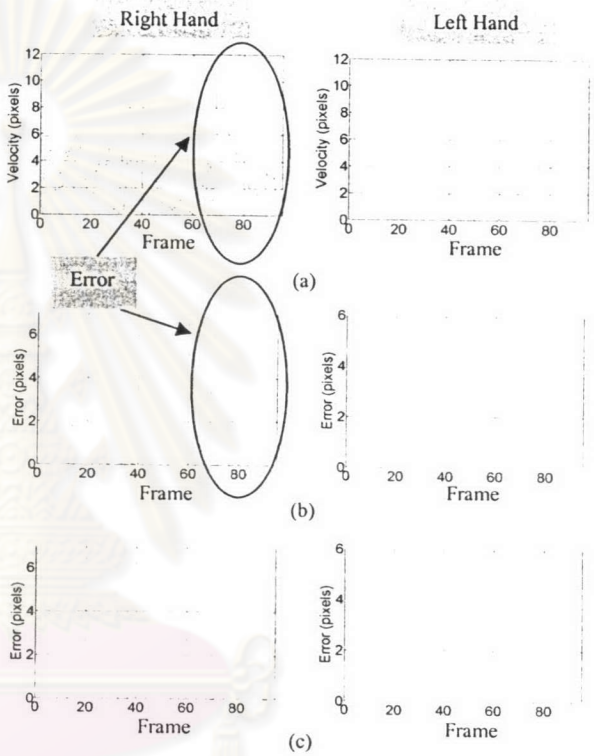
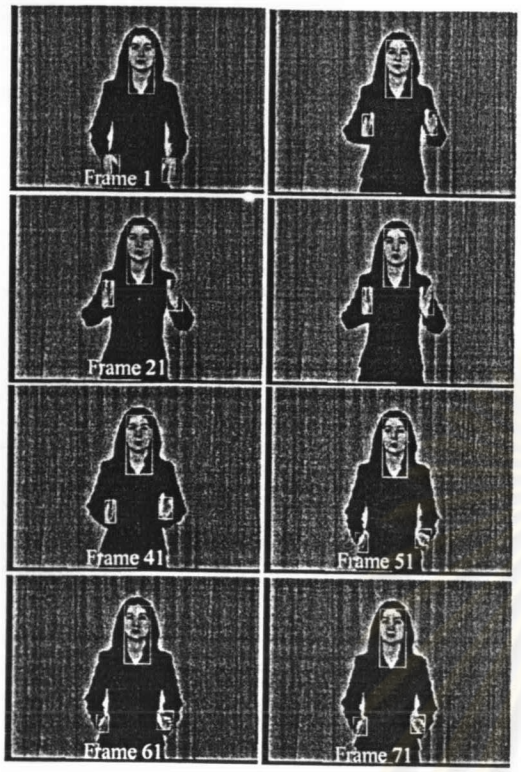


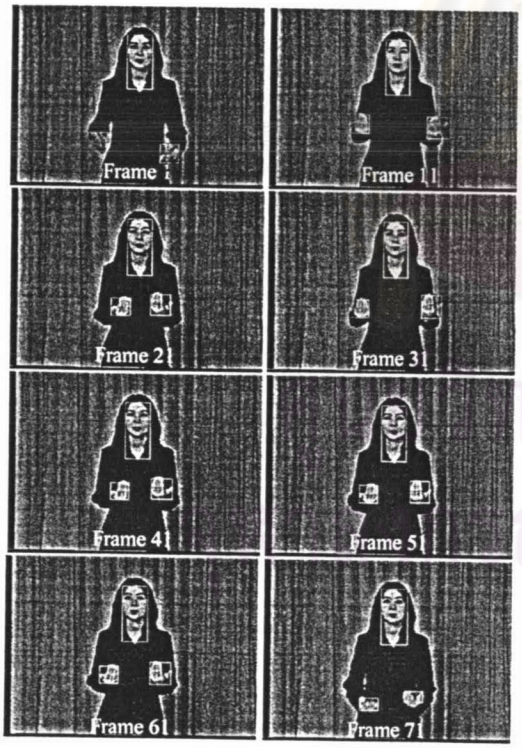
Figure 8. Euclidean error of each frame at seq.2
(a) Velocity (b) Position Error eq. (8)
(c) Position Error (Proposed)

5. Conclusions and Future Work

We have presented the face and hand detection and tracking for sign language recognition. Skin color feature is selected in detection algorithm. After face and hand blobs has been located separately by using size and facial feature. In tracking part, face is easily to compute the future position but location of hands are calculated the motion estimation model which are proposed it has slightly position error. Simulation results indicate our proposed algorithm can track face and hand with greater precision with negligible computational complexity increase.



(a) Sequence 1



(b) Sequence 2

Figure 9. Tracking result

6. Acknowledgements

This work was supported in part by the Cooperation project between Department of Electrical Engineering and private sector for research and development, Chulalongkorn University, Thailand.

7. References

[1] N. Tanibata, N. Shimada and Y. Shirai, "Extraction of Hand Features for Recognition of Sign Language Words," in *Proc. of Int. Conf on Vision Interface*, pp.391-398, 2002.

[2] A. Shamaie and A. Sutherland, "A dynamic model for real-time tracking of hands in bimanual movements," in *5th International Gesture Workshop, Geneva*, April 2003.

[3] K. Imagawa, S. Lu and S. Igi, "Color-based Hands Tracking System for Sign Language Recognition," *Face and Gesture (FG 1998)*, pp.462-467, 1998.

[4] Jong-Sung Kim; Won Jang; Zeungnam Bien, "A dynamic gesture recognition system for the Korean sign language (KSL)," *IEEE Trans. on Systems, Man and Cybernetics*, vol.26, pp. 354 - 359, April 1996.

[5] N. Soontranon, S. Aramvith and T.H. Chalidabhongse, "Face and Hand Localization and Tracking for Sign Language Recognition," in *Proc. International Symposium on Communication and Information Technologies (ISCIT'04)*, October 2004.

[6] C. R. Wren, A. Azarbayejani, T. Darrel, and A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp.780-785, July 1997.

[7] R. L. Hsu, M. Abdel-Mottaleb and A. K. Jain, "Face Detection in Color Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.2, No.5, pp.696-706, May 2002.

[8] J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink and G. Sagerer, "Improving Adaptive Skin Color Segmentation by Incorporating Results from Face Detection," *IEEE Int. Workshop on Robot and Human Interactive Communication*, September 2002.

[9] M. J. Jones and J. M. Regh, "Statistical color models with application to skin detection," in *Proc. Computer Vision and Pattern Recognition*, pp.274-280, June 1999.

[10] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Reading, MA: Addision-Wesley, 1992.

[11] Thai sign language video sequences, Ministry of Education.

ประวัติผู้เขียนวิทยานิพนธ์

นายนรุตม์ สุนทรานนท์ เกิดเมื่อวันที่ 22 พฤศจิกายน 2521 สำเร็จการศึกษาปริญญาวิศวกรรมศาสตรบัณฑิต ภาควิชาวิศวกรรมโทรคมนาคม คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ในปีการศึกษา 2544 และเข้าศึกษาต่อในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต ในสังกัดห้องปฏิบัติการกรรมวิธีสัญญาณดิจิทัล ภาควิชาวิศวกรรมไฟฟ้า ที่จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2545