Color-Based Hands Tracking System for Sign Language Recognition

Kazuyuki Imagawa
Matsushita Electric Industrial Co.,Ltd.
693-47, Kawazu, Iizuka,
Fukuoka 820-0067, Japan
imagawa@qrl.mei.co.jp

Shan Lu, Seiji Igi
Communications Research Laboratory,
4-2-1, Nukui-Kitamachi, Koganei,
Tokyo 184-0015, Japan
lu,igi@crl.go.jp

Abstract

This paper describes a real-time system which tracks the uncovered/unmarked hands of a person performing sign language. It extracts the face and hand regions using their skin colors, computes blobs and then tracks the location of each hand using a Kalman Filter.

The system has been tested for hand tracking using actual sign-language motion by native signers. The experimental results indicate that the system is capable of tracking hands even while they are overlapping the face.

1. Introduction

We are developing a bi-directional translation system between Japanese Sign Language (JSL) and Japanese in order to help the hearing impaired communicate with normal speaking people through sign language [1]. Since sign language is the primary communication method used in the daily life of the hearing impaired, we think that it is appropriate to use it as a means of providing information from and interacting with a computer.

Sign language is a set of sophisticated hand gestures using both hands. It is known to be composed of basic units which consist of hand shape, movement, and position [2, 3]. Therefore, one must track both hands in order to detect these units for sign language recognition.

To date, there have been several sign-language recognition systems using the vision-based approach [4, 5, 6]. These systems track hands during sign language to recognize the action. Although these systems demonstrated skin color hand tracking, some of them are limited to single-handed sign languages. The rest were not evaluated with real sign language motion by native signers. We also develop a sign-language recognition system for native signers[7]. However, this system requires the user to wear colored gloves to enable the system to track hands because the system needs to distinguish between a hand region and a face region while the hand is overlapping the face. In sign language, people often move their hands in front of their face. Therefore, tracking hands during this overlapping duration is a major issue in the vision-based approach.

In this paper, we present a real-time hand-tracking system for sign language recognition. The system extracts the face and hand regions using their skin colors, computes blobs, and then tracks the location of each hand using a Kalman Filter. In order to track hands while they overlap the face, we focus on the fact that the face does not move as much as the hand during sign-language motion. Therefore, our system uses not only blobs in a single frame but also motion difference blobs. In order to be able to compute the motion difference blobs stably while the hand is overlapping the face, we present a new color-extraction method based on *histogram backprojection*[8]. This method enables only the hand and face regions to be extracted from complex backgrounds. It also enhances the contrast of these regions. This enhancement enables the system to compute motion difference blobs stably while the hand is overlapping the face.

In order to evaluate our system, we prepared not only images which are taken in our laboratory but also ones in the published VHS video library [9], which contains many interviews with native signers performing sign language at natural speed. In this paper, we present our experimental results using this library, and discuss the limitations of our system.

2. Defining the Scene

Before we present our system overview, we define the sign language scene which our system expects. We as-

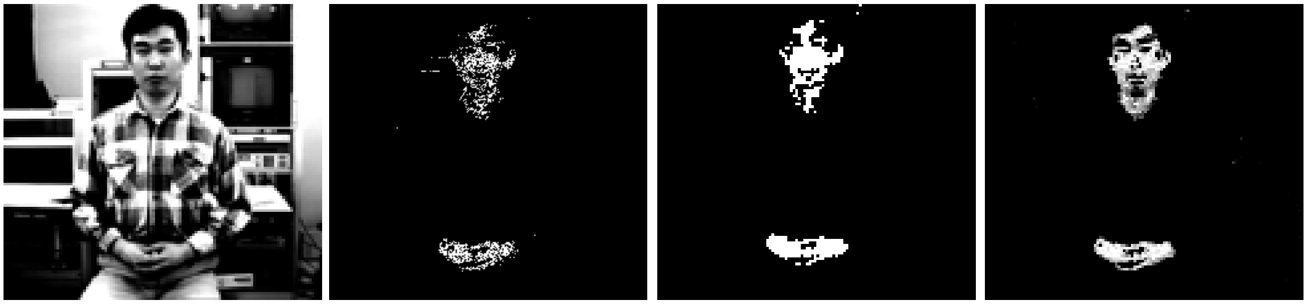
(a) Original image (b) Skin segmentation (c) Skin color regions (d) Extracted image

Figure 1. Extracting face and hand regions

sume that the scene for recognizing sign language is relatively static and that sign-language motion is acquired with a stationary camera. The person performing sign language is required to wear long-sleeved clothes which are a different color from the skin in order to focus on the influence of the head upon the tracking hands.

3. Extraction of Hand and Face Regions

Our system extracts hand and face regions using the color distribution from the sign-language image sequence. For real-time processing, we prepare a three-dimensional look-up table (3-D LUT) which is used for setting the color distribution. In the first frame of the image sequence, the system sets the 3-D LUT according to the face, hands, and other colors of the frame since we assume that the illumination condition does not change while the person is performing sign language. Then, the system extracts the face and hand regions from all frames of the sequence using the 3-D LUT.

3.1. Automatic Skin Segmentation

In the first frame of the sequence, the system segments either skin color regions or a background region. In order to segment the skin color regions, the system must locate the face and hand regions. The skin colors of different people are very close in the chromatic color plane but they differ in intensities. Many papers have shown that a face can be located by disregarding the luminance components of the color space [10, 11]. We also locate the face and hand regions by using the chromatic color distribution of skin color.

First, we obtain multiple samples of skin from images of several individuals under varying illumination conditions. According to the color of each pixel, a hue-saturation map for the skin color is made. We apply this map to the first frame of the sequence and segment skin color regions. The skin color regions are dilated and eroded. Then, the connected regions, which have fewer pixels than a threshold value determined by the system, are deleted in order to eliminate the influence of noise. An example of this step is shown in Fig. 1(a), (b), and (c).

3.2. Improved Histogram Backprojection

As described before, our aim is to get the system to track the hand when it overlaps the face. Our method not only needs to extract skin-color regions but also to enhance the contrast of the extracted image in order to detect hand motion when the hand overlaps the face. Therefore, we present a new method, which is based on *histogram backprojection* [8]. We then set the 3-D LUT with this method by using the skin-color regions and the background region.

The overview of our method is described below. Due to the skin segmentation, all pixels in the first frame of the sequence can be classified into either the skin color regions M or the background region B. According to the multi-dimensional color histogram of both regions, a combined histogram C is defined as

$$C_j = \max\left(\frac{w_M M_j - w_B B_j}{C_{max}}, 0\right) \times D,$$

where j is the index of each histogram bin, C_{max} is the maximum value of the combined histogram, D is the range of the output image, e.g., 255 for an 8-bit image, and w_M and w_B are weighted values representing the sensitivity for C in each histogram, which are implemented as $w_M = w_B = 1$ in our system. In order to

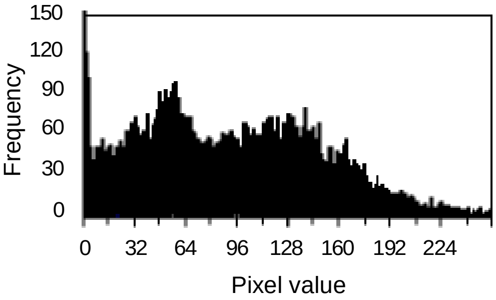

Figure 2. Histogram of the extracted regions

implement the method in our system, we prepared a small 3-D LUT memory for an 8 bits × 3 RGB color space. Each axis is divided into widths d, which is defined as 32 in our system. Here, the 3-D LUT has 512 ($2^3 \times 2^3 \times 2^3$) unit cubes. The combined histogram C is set to the 3-D LUT.

All pixels in the image are converted using the 3-D LUT as follows. If the color value of the pixel is the same as the mean color value of a unit cube, the value of the pixel is converted to the value of the unit cube. If the color pixel value is a value other than the mean color value of a unit cube, the pixel value is converted to a value which is interpolated by the $PRISM$ algorithm [12] using the values of six neighboring cubes. The extracted image is shown in Fig. 1(d) and the histogram of the extracted regions is shown in Fig. 2. Our method is performed to highlight the skin areas and thereby accentuate intensity variations in these parts of the image. These emphasized intensity variations enable the system to detect hand motion clearly while the hand overlaps the face.

4. Tracking Hands

After extracting hand and face regions from a sign-language image sequence, the system computes blobs using the extracted image sequence and then tracks the right and left hands. In order to track hands when they overlap the face, blobs are computed not only for a single frame image but also for a time differential image between continuous frames. First, the single frame image and the time differential image are binarized. Next, connected regions that have more pixels than a threshold determined by the system are found in these images, and labeled as either a *static* or *motion blobs*. Fig. 3 shows a typical example of static blobs and motion blobs.

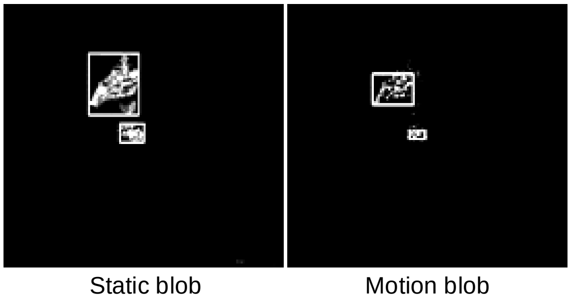

Figure 3. Static blobs and motion blobs. Each blob is marked by a white bounding box. Although the face and hand regions appear as one static blob, motion blob by the hands are appeared

After computing blobs, the system assigns one of them to each hand. The assignment is done by predicting the location of each hand using a Kalman filter [13]. Then, the nearest blob to the predicted location within the threshold value is assigned to either the left or right hand and is labeled as *“hand blob”*. The threshold value is determined by the length of the palm because a motion blob may be generated by the edge of the hand. Thus, the system determines hand blobs for both hands. After each blob has been labeled, the system updates the observed location of each hand using the hand blobs. The observed location is updated by the center of the hand blob. If a hand blob exists in the current frame, the observed location is updated using the current hand blob. However, if the hand stops in front of the face, no hand blobs originating from the hand appear. Here, the observed location is updated from the location of the last hand blob because this situation is caused by the hand stopping in front of the face.

In order to apply a Kalman filter, we assume that the movement of the hand is sufficiently small during a step interval ΔT. Accordingly, a dynamic process can be used to describe the x or y coordinate of the center of the hand on the image plane with the state vector $\mathbf{x}$ which includes its position and velocity. The dynamic process is defined as

$$\mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{G}_k w_k,$$

where

$$\mathbf{F}_k = \begin{pmatrix} 1 & \Delta T \\ 0 & 1 \end{pmatrix}, \; \mathbf{G}_k = \begin{pmatrix} \Delta T^2/2 \\ \Delta T \end{pmatrix}.$$

The system noise is modeled by w_k, an unknown scalar acceleration. An observation model is given by

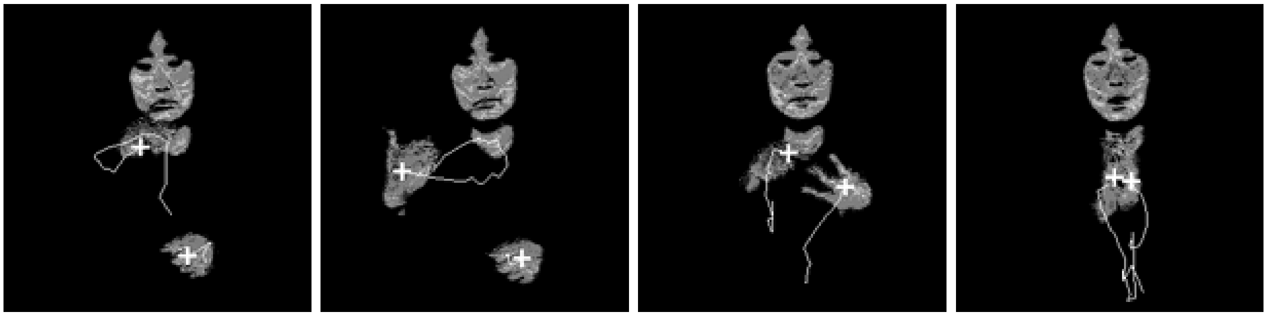

Figure 4. Examples of resulting trajectories in this experiment. The tracking result is represented as a cross cursor overlaid on the extracted image. The hand trajectories from the past 30 steps are drawn by the white solid lines

$$z_{k+1} = \mathbf{H}\mathbf{x}_{k+1} + v_k,$$

where $\mathbf{H} = [1, 0]$, $\mathbf{x}_{k+1}$ is the actual state vector at time $k+1$, v is measurement noise, and $\mathbf{z}_{k+1}$ is the observed location at time $k+1$. The noise covariances are determined by experiments so that the system can perform optimal tracking.

5. Implimentation

The system has been implemented on a standard PC-AT, with a custom daughter board which provides the color extraction by the 3-D LUT at full frame rate for extracting the face and hand regions, and MVC 150/40 image processing boards supplied by Imaging Technology Inc in order to track hands. The image processing boards are used for real-time processing. The boards input the image sequence, in which the hand and face regions are extracted by the color extraction board, and digitize one frame with an image area of 512 × 480 pixels. They compute the motion difference image and extract features of streaks in both the single frame and motion difference image at full frame rate (30 Hz). Blob features are computed using these streaks, and computing blobs and hand tracking are implemented on the PC-AT. The total computer processing time for one frame was approximately 60ms. Thus proposed methods can be employed for real-time applications.

6. Experiments

The system was tested on sign language performed by native signers. The following sign-language image sequences, which are acquired under the definition described in Section 2, were used for this experiment.

- Lab sequences: the image sequences of two native signers performing sign language were taken in our laboratory.
- Published sequences: the image sequences of three native signers performing sign language were chosen from the VHS video library [9].

In each image sequence, the experiment used sign language motion for approximately 30 seconds.

6.1. Experimental Results

Fig. 4 shows examples of resulting trajectories in this experiment. These trajectories include duration when the hand overlaps the skin color region, neck. Nevertheless, the trajectories are smooth and the system succeeded in tracking the hands. This is a result of detecting motion blobs in the overlapping duration.

In order to analyze the tracking results, we counted the number of frames in which the cross cursors overlaying the image were within the hand regions and then classified these frames as either overlapping periods or not. An overlapping period means the duration when either one or both hands overlap the skin color region. In order to classify these frames, we then checked whether the hand region and the face region appeared as one static blob or not.

Table 1 lists the results of this experiment. We obtained a success rate of 86% for the tracking for all sequences, 83% for overlapping periods, and 90% when the hand did not cross the face. The success rates of the published sequences C, D and E were lower than those

Table 1. Results of the experiment. Sequences A and B are lab sequences, and C,D and E are published sequences.

	Total		Overlapping period		Non-overlapping period	
Seq.	S/T (frames)	R (%)	S/T (frames)	R (%)	S/T (frames)	R (%)
A	859/900	95	526/558	94	333/342	97
B	859/915	94	120/146	82	739/769	96
C	775/914	85	470/589	80	305/325	94
D	724/902	80	524/638	82	200/264	76
E	642/866	74	413/556	74	229/310	74
Total	3859/4497	86	2053/2487	83	1806/2010	90

S : success frames
T : total frames
R : success tracking rate

Table 2. Results of error analysis. For each sequence, the number of frames in which the system made an error is indicated. The errors were classified manually

	Overlapping period				Non-overlapping period			
Seq.	TH	BH	TA	OT	TH	BH	TA	OT
A	4	14	21	2	1	8	21	2
B	46	0	10	0	16	0	10	0
C	27	55	57	0	22	40	56	1
D	81	52	22	23	54	20	22	18
E	123	61	21	19	103	13	13	14
Total	281	182	131	44	186	81	122	35

TH : Hand blob is trapped by the head.
BH : Both hand regions are merged.
TA : Hand blob is trapped by the another hand.
NH : Other reasons.

for the lab sequences A and B. In fact, the subjects for lab sequences performed sign language more slowly and clearly than the subjects in the published sequences so that the non-native people could understand their motion. In the published sequences, the subjects did sign language more naturally. They used their eyebrows, mouth, and head orientation to help convey meaning. In addition, they used facial expressions, and head and body movements in order to express emotions. These movements caused a reduction in the success rate in the overlapping period.

6.2. Error Analysis

In order to investigate the limitation of the system, we analyzed the tracking error. We checked the number of frames in which the system made an error. Table 2 lists the results of the error analysis. These errors can be roughly classified into four categories.

The three major categories -TH, BH and TA - are caused by insufficient distinction between blobs which are generated by two hands and a head. For example, the most erroneous case TH occurred when the hand blob was trapped by motion blobs generated by the moving head in the overlapping period. For example, the signers of sequence D and E sometimes nodded during sign language motion. Here, the system needed to select the hand blob from motion blobs generated by both hand and head in the overlapping period. In particular, the system had to ignore motion blobs surrounding the predicted location when the hand stopped in front of the face and the head moved. It was impossible for the system to ignore these motion blobs because the proposed algorithm does not at present address the issue of compensating for head movements.

The second erroneous case BH occurred when both hand regions merged or one hand occluded the other hand, and then appeared as a single static blob. Sometimes hands cross during a sign so that the left hand is distinctly on the right side. Also, hands collided and bounced back. This was a problem for hand tracking by using only the Kalman filter. The system did not have information for indexing of left/right hand.

Although our system could distinguish between a hand and the static head, it could not completely classify blobs into three objects (head, left hand and right hand). In order to overcome these problems, we need to develop a 3-D human model for sign language recognition. Then, the system will acquire the sign language with multiple cameras in order to deal with occlusion of the hands. 3-D model-based techniques have been presented in [14, 15]. These techniques, however, do not seem to classify blobs into these objects completely. In the current method, we used empirical knowledge that the head does not move as much as the hands during sign language while the hand is overlapping the face. Similarly, we will apply extensive knowledge of sign language to the 3-D model and estimate the 3-D motion of the subject according to constraints resulting from this knowledge.

Other erroneous cases might have occurred due to insufficient coordination of the threshold value when the system computed blobs. For example, a computed blobs was too small to be regarded as blobs originating from the hand because its area was less than the threshold value of the system. Consequently, no hand blob appeared in the current frame, and the tracking result stayed in the previous location. This error can be reduced by eliminating noise and by appropriately setting the threshold value.

7. Conclusion and Future Work

The problem of hand tracking when the hand overlaps the face has been addressed in this paper. This is a prerequisite for most methods of tracking hands in sign language. This paper presented a new hand tracking system for sign language. Experimental results show that the system is capable of tracking hands in real sign-language motion, although it can not track hands completely since it has problems distinguishing between multiple objects - the head, left hand and right hand.

Currently, our system is limited to a static head and it cannot detect hand shape. However, hand shape is necessary even in front of the face. In addition, in conversational sign language, the eyebrows, mouth, and head orientation are used to help convey meaning. In the future, our system will be extended to incorporate a 3-D human model with a multi-view approach. Then, the system will be integrated with a feature extractor for sign-language recognition system including even hand shape, and mouth and head orientation.

Acknowledgments

We thank the members of the Deaf Association of Tokorozawa in Japan for their cooperation in the experiment.

References

[1] S. Lu, S. Igi, H. Matsuo and Y. Nagashima. "Towards a Dialogue System Based on Recognition and Synthesis of Japanese Sign Language", *Proc. Bielefeld Gesture Workshop 1997*,1997.

[2] W. Stokoe, D. C. Casterline and C. G. Groneberg. "A Dictionary of American Sign Language on Linguistic Principles," *Linstok Press*,London,1976.

[3] K. Kanda. "A Computer Dictionary of Japanese Sign Language: General Idea and its Present Stage," *The fifth Int'l Symposium of Sign Language Research*,1992.

[4] S. Tamura and S. Kawasaki. "Recognition of sign-language motion images," *Pattern Recognition*, Vol. 21,No. 4,pp. 343-353,1988.

[5] C. Charayaphan and A. E. Marble. "Image-processing system for interpreting motion in American Sign Language," *J. of Biomedical Engineering*, Vol. 14,pp. 419-425,1992.

[6] T. Starner and A. Pentland. "Real-Time American Sign Language Recognition from Video Using Hidden Markov Models," *Proc. Int'l Symposium on Computer Vision*, pp. 265-270,1995.

[7] H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata and T. Teshima. "Recognition Algorithm with Non-contact for Japanese Sign-Language using Morphological Analysis," *Bielefeld Gesture Workshop 1997*, 1997.

[8] M. J. Swain and D. H. Ballard. "Color Indexing," *Int. J. Comput. Vision*, Vol. 7,No. 1,pp. 11-32,1991.

[9] Sign Factory eds.,"Sign Language Journal," *Sign Factory*,Japan,Spring,1996.

[10] J. Yang and A. Waibel. "A Real-Time Face Tracker," *Proc. WACV'96*, pp. 142-147,1996.

[11] K. Sobottka and I. Pitas. "Segmentation and Tracking of Faces in Color Images," *Proc. 2nd Int'l Conf. on Automatic Face and Gesture Recognition*, pp. 236-241,1996.

[12] K. Kanamori, H. Kotera, O. Yamada, H. Motomura, R. Iikawa and T. Fumoto. "Fast color processor with programmable interpolation by small memory (PRISM)," *J. of Electronic Imaging*, Vol. 2,No. 3, pp. 213-224,1993.

[13] R. E. Kalman. "A new approach to linear filtering and prediction problems," *Trans. ASME, J. Basic Eng.*, Vol. 82D,No. 1,pp. 35-45,1960.

[14] C. R. Wren, A. Azarbayejani, T. Darrel and A. P. Pentland. "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. PAMI-19,No. 7,pp. 780-785,1997.

[15] D. M. Gavrila and L. S. Davis. "3-D model-based tracking of humans in action: a multi-view approach," *Proc. CVPR'96*, pp. 73-80,1996.