Segmenting Hands of Arbitrary Color

Xiaojin Zhu Jie Yang Alex Waibel
Interactive Systems Laboratories
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{zhuxj, yang+, ahw}@cs.cmu.edu

Abstract

Color has been widely used for hand segmentation. However, many approaches rely on predefined skin color models. It is very difficult to predefine a color model in a mobile application where the light condition may change dramatically over time. In this paper, we propose a novel statistical approach to hand segmentation based on Bayes decision theory. The proposed method requires no predefined skin color model. Instead it generates a hand color model and a background color model for a given image, and uses these models to classify each pixel in the image as either a hand pixel or a background pixel. Models are generated using a Gaussian mixture model with the restricted EM algorithm. Our method is capable of segmenting hands of arbitrary color in a complex scene. It performs well even when there is a significant overlap between hand and background colors, or when the user wears gloves. We show that the Bayes decision method is superior to a commonly used method by comparing their upper bound performance. Experimental results demonstrate the feasibility of the proposed method.

1. Introduction

Hand segmentation is a prerequisite for many gesture recognition tasks [1]. A popular feature for hand segmentation is skin color. Many approaches rely on predefined skin color models, which work well in constrained environments. In this paper, we investigate a hand segmentation problem for a wearable computer application. We show that it is very difficult to predefine a color model in this case because the light condition may change dramatically over time. To solve the problem, we propose a novel statistical approach to hand segmentation based on Bayes decision theory. The new method still relies on color information, but requires no predefined skin color model. The key innovation is the dynamic construction of hand and background color models using Gaussian mixture models and the restricted EM algorithm. The proposed method is capable of segmenting hands of arbitrary color in a complex scene. Another contribution of this paper is the study of upper bound performance for color-based hand segmentation. The study indicates the limit for the color-based approaches. We demonstrate that the proposed method is superior to a commonly used threshold method.

The rest of the paper is organized as follows. Section 2 describes a motivating application where hand segmentation approaches based on predefined color models are challenged. Section 3 presents the new method and algorithms. Section 4 investigates the performance of the new method, and compares it to the threshold method. Section 5 addresses the limitations of the proposed method and possible improvements.

2. A Motivating Application

Recent technological advances have made wearable computers available for many different applications. However, how to efficiently interact with a wearable computer is still an open question. A gesture interface is certainly a good solution to the problem. Finger can been used as pointers for menu selection [2]. In a wearable environment a head-mounted see-through display may jitter due to involuntarily head motion. In this case it may be hard to point a finger at a menu item, because the item is displayed at a fixed position on the head-mounted display and moves around with the head.

We propose a new menu selection paradigm for wearable computers, namely the "finger menu". It works on a wearable computer with a head-mounted see-through display and a head-mounted video camera. Unlike traditional

GUI's that display menu items at fixed positions on the screen, our method associates menu items onto the user's five fingers. It works as follows:

1. The user sees through the head-mounted screen. The head-mount camera also monitors the same scene. Since the user's hand is not in the scene, no menu is displayed. (Figure 1a).

2. The menu system is activated when the user moves his hand, widely opened, into the scene. The system detects the hand with the camera, and displays five menu items at appropriate positions so that they appear on the fingertips through the head-mount display. (Figure 1b).

3. The menu items 'float'. When the hand moves, they move accordingly so that they stay on the fingertips. Thus there is no need for the user to move the hand to a specific place to make a menu selection. (Figure 1c)

4. By bending a finger as if it is a virtual 'click', the user can select the menu item on that fingertip (Figure 1d).

5. The menu system is de-activated when the user moves his hand out of the view.

The advantages of this paradigm include: intuitive interaction, a user can operate the device with little or no training; efficient operation, there is no need to move a finger to a specific place to make a selection, which could be a hard head-hand coordination task; and no need for special pointing hardware.

In order to implement the finger menu system, we need to recognize hands from images taken by the head-mounted camera. Hand segmentation is an essential preprocessing step. However it is a hard problem for the following reasons:

- There are no restrictions on the background.

  The camera moves with the user's head.

  The light conditions may change dramatically. This includes changing shadow and varying light colors, e.g. under a sodium-vapor lamp.

There are numerous publications on hand segmentation. Two common methods are background subtraction and skin color segmentation. Obviously background subtraction is infeasible since there is no constant background. Color segmentation [3] [4] [5] [6] [7] [8] [9] [10] is more suitable in our case. Nevertheless, previous methods often use one static skin color model, which is inadequate for us. In the rest of this paper, we present a new way of segmenting hands with color information.

We formulate the hand segmentation problem as follows: The hand is known to be in an image. The hand color is unknown in advance (different environments may result in different hand colors), but is assumed to be largely consistent within the image. In addition, we are concerned with initial hand segmentation, not subsequent hand tracking. Thus we limit ourselves to a single image. Under these conditions, we want to segment the hand from the background, i.e. for each pixel in the image, we want to classify it as either a hand pixel or a background pixel.

To facilitate the discussion, we introduce our hand image data set. A user recorded image sequences of his hand with a head-mounted camera while performing the various finger menu actions. The actions were performed at typical places where a wearable computer is used including: office, home, vehicles, parks, and streets etc., with various backgrounds and light conditions. Some sequences were taken while the user was wearing gloves, which is a reasonable situation for working in the field. From the sequences, 326 images were selected randomly and the hands in these images were manually segmented. Each image is 80 * 60, 24 bit color. The 326 images were randomly divided into two halves, with 163 images as training data and the other 163 images as test data. Figure 2 shows random samples of the training data set.

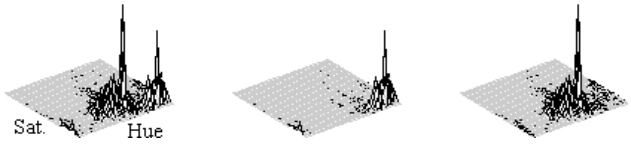
In this work we use the HSV color space [11] instead of the RGB color space. Moreover, we use only Hue and

Saturation and ignore V (brightness) in order to minimize the influence of shadow and uneven lighting. We plot Hue-Saturation histograms for analysis. Since the images are manually segmented, we are able to plot the color histograms for the overall image, the hand portion, and the background portion respectively, as shown in Figure 3. Two observations arise after investigating some images:

- The peak of hand color is not fixed at a certain position on the H-S plane (compare Figure 3b with 3d). This means we cannot build a static hand color model for all images. Instead, we will build a different hand color model for different image.

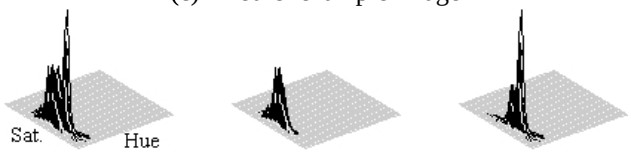
  The hand color may partially overlap with the background color, as shown in Figure 3d. This means some hand pixels and background pixels have the same color. Thus misclassification is inevitable in color segmentation. However we want to minimize the error.

(a) An example image, its hand and background portion



(b) Overall, hand and background histograms of (a)

(c) Another example image



(d)H istograms of(c)

3. A Novel Color Segmentation Method

In this section, we first introduce a Bayes decision theory framework for segmentation. The framework needs a hand color model and a background color model to work. We then present an approach to build the models dynamically for any given image using Gaussian mixture models and the restricted EM algorithm, which is the key innovation of this paper.

Given the color and coordinates of a pixel, we want to classify it as a hand pixel if

(1)

Applying the conditional version of Bayes rule, we get

We assume is conditionally independent of x, y given hand, i.e. . Thus

And similarly for . Note 1, therefore (1) becomes

1 (2)

(2) is our Bayes decision criterion, which will be used to classify the pixels in an image.

We need three models to compute (2). The first one, , is the hand color model of an image. The second one, , is the background color model of the image. These two models need to be built for each image dynamically, as discussed in Section 3.2. The third one, , describes the spatial distribution of hand pixels, i.e. how likely the pixel is a hand pixel. We can estimate it from the training data set as follows:

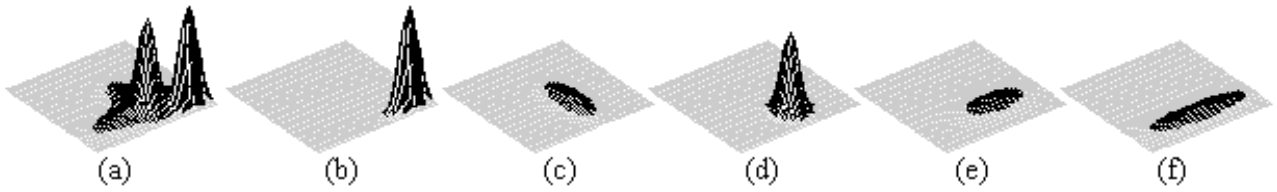
(3)

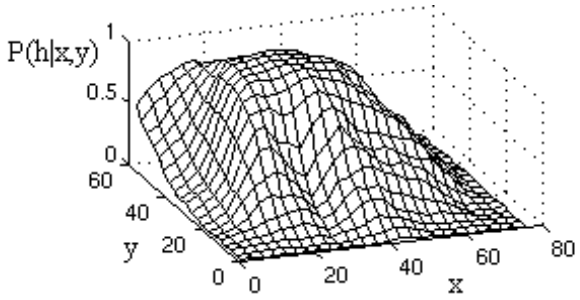
1

where 1 if is a hand pixel in image , 0 otherwise. Figure 4 is the distribution of our training data set. The highest region corresponds to the palm. Since the user tends to place the hand at the center of the view, this distribution is reasonable.

We need to estimate the hand color model and the background color model for any given image. Since hand color may change from image to image, and the hand color may partly overlap with the background color (as in Figure 3), this is not a trivial task.

One observation is that hand color is largely consistent within an image. This means hand pixels tend to concentrate





and form a single peak in a hand color histogram. Intuitively it suggests to model the hand color with a Gaussian distribution. This leads to the following method.

Given an image, we can easily compute its overall color distribution by normalizing the color histogram of the whole image:

It has the following relationship with the (yet unknown) hand color model and background color model:

$$+ \qquad (4)$$

$$1$$

where is the percentage of hand pixels in the image, or the relative hand size (not to be confused with in the previous section, which is a pixel level value).

We can approximate with a Gaussian Mixture Model (GMM) [12]. The GMM is a weighted sum of Gaussian distributions :

$$\sum \qquad 1$$

is determined empirically. We found 5 is sufficient. Let , Σ denote 's mean and covariance matrix respectively. With the Expectation Maximization (EM) algorithm [13] [14], we can train the GMM, that is to find a set of parameters 's, 's and Σ's, such that . Figure 5 is the GMM trained with random starting parameters for the image in Figure 3a. Note how well the GMM approximates the actual overall color distribution in Figure 3b. Also note how well (Figure 5b) resembles the actual hand color distribution in Figure 3b. Since we assume the hand color can be modeled with a Gaussian distribution, it is natural to think can be used as the hand color model. An immediate question follows: Can we guarantee that approximates the actual hand color distribution well enough for any image? If the answer is yes, then by the definition of GMM:

$$\qquad (5)$$

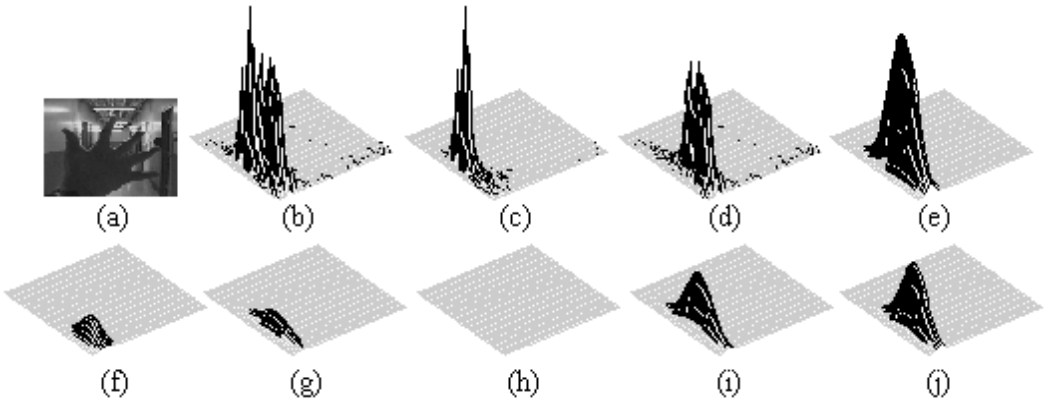
And comparing (4) with (5), we would have the following parametric forms to solve the problem:

$$\qquad (6)$$

$$\qquad (7)$$

Unfortunately, in most cases the answer is no if we use the standard EM algorithm. In fact, the standard EM algorithm only guarantees the overall fitting of the whole distribution, but has virtually no control over the individual component Gaussians. There is no guarantee that any of the Gaussian components will be a reasonable approximation to the actual hand color distribution. Better starting parameter heuristics will not help either. Figures 6(a-d) show an example image and its histograms. Figures 6(f-j) are the Gaussian components obtained using the standard EM algorithm. Obviously none of the Gaussian components resembles the actual hand color distribution (Figure 6c).

However, we will show that with certain modifications to the standard EM algorithm, we can enforce , the first Gaussian component of a GMM, to be a good approximation of the hand color distribution such that we can use (6)(7). This is the key innovation of our method.



During the Maximization step of the standard EM algorithm, the mean, covariance and weight of each Gaussian component can be adjusted freely. However we can fix some parameters to certain values, or limit their ranges during EM training. We call it the restricted EM algorithm. It will still converge to a local maximum in terms of likelihood [15]. More specifically, in this paper we will fix $\mu_1 = \tilde{\mu}_1$ and limit Σ_1 to be within range . The meaning and value of $\tilde{\mu}_1$, , and will be discussed in next section. By restricting these two parameters, we can enforce N_1 to approximate the hand color distribution. The restricted EM algorithm is:

Initialization:
Let $_1$ $\tilde{\mu}_1$, $_1$ 2.
Set other parameters randomly.
During the E-step:
Collect counts as in standard EM algorithm.
During the M-step:
1. Adjust only $_2$ as in standard EM, leave $_1$ unchanged.
2. Adjust Σ_1 Σ as in standard EM
3. Adjust $_1$ as in standard EM
4. If $_1 >$ Then
$*$ $_1$ 1 $\frac{1}{1 \quad 1}$ 2
5. Similarly for the $_1 <$ case.
Iterating the E-step and M-step until converge.

Figure 7 shows the effect of the restricted EM algorithm. Figure 7a is the GMM obtained with the restricted EM algorithm on the same image of Figure 6a. Note the first Gaussian component (Figure 7b) now approximates the actual hand color histogram (Figure 6c).

$\tilde{\mu}_1$ is the estimated mean of the hand color distribution. It needs to be estimated for each image as follows. Consider three random variables , , and which take value of possible colors. has distribution , and has distribution . is the color of pixel . We assume the following generative random process:

1 _

That is, pixel 's color is generated in such a way: Firstly the 'identity' of this pixel is chosen to be 'hand' with probability and 'background' with probability 1 (see Figure 4 for the distribution). Secondly, if it is 'hand' a color is randomly picked for the pixel according to , otherwise the color is picked according to .

Now consider a set of pixels in an image with hand probability , where is a fixed value between 0 and 1. Denote this set as . The expectation of the previous equation over this set is

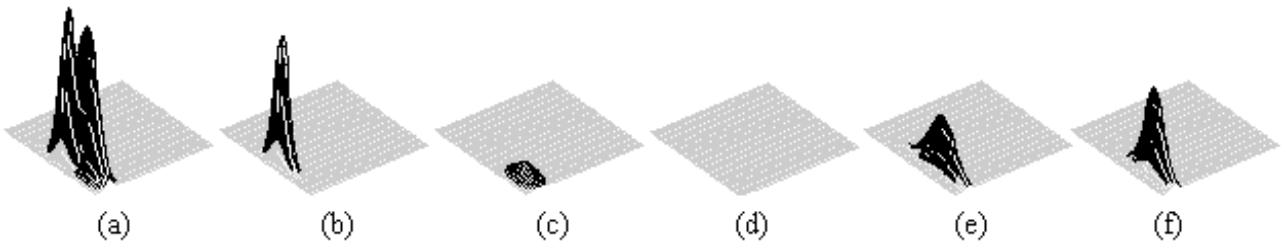
$\in$: E
1

By definition $\tilde{\mu}_1$. And for :

$\sum$ $\frac{}{}$

Therefore we get

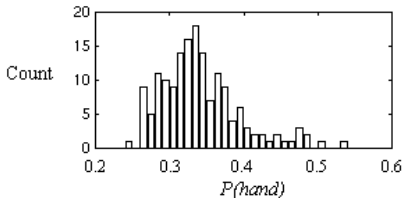
$\tilde{\mu}_1$ $\frac{1}{}$



$$\frac{1}{} \quad (8)$$

where can be estimated from the training data. In particular, if we can find some close to 1 such that is large enough, we can use the approximation

The restrictions , control $_1$, which can be interpreted as , the relative hand size. They are estimated from the training data. The distribution of in the training data is plotted in Figure 8. We compute the



8

mean and standard deviation of this distribution. Since we expect the hand size in a new image to be comparable to those in the training data, we let

$$= \quad 2 + \quad 2 \quad (9)$$

The complete algorithm is as follows.

During training:
- Build with (3)
- Estimate the weight restrictions , with (9)

During decoding:
- Estimate the mean restriction $\tilde{}_1$ with (8)
- Run the restricted EM algorithm in Section 3.3
- Generate the hand color model with (6)
- Generate the background color model with (7)
- Classify each pixel with (2)

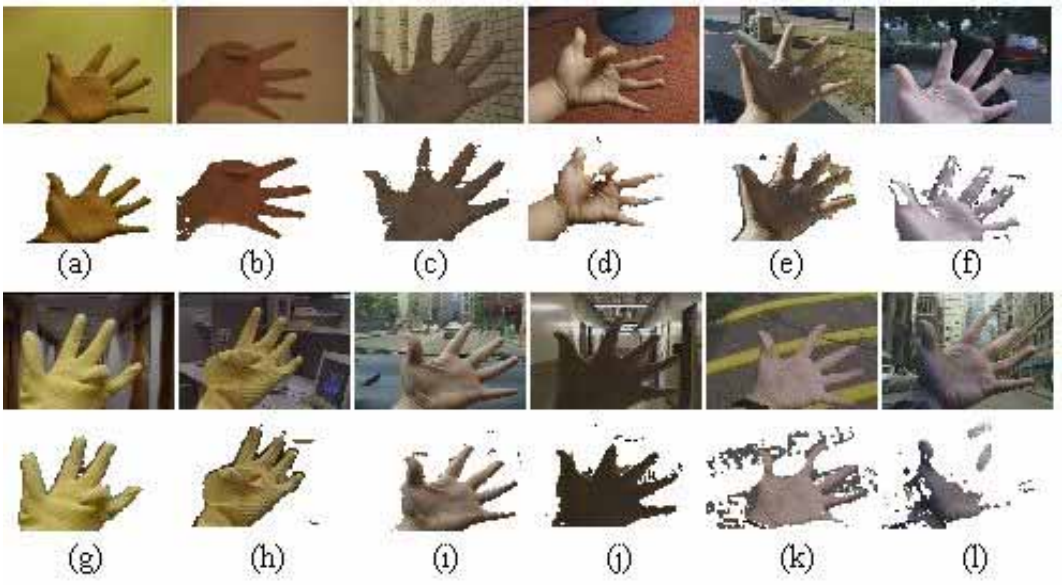
4. Performance Analysis

We tested the proposed method on the 163 test images. We achieved an average false positive (background misclassified as hand) rate of 4.0%, false negative rate of 7.5%, and total error rate of 11.5%. The decoding takes about 0.1 second for each image on a PC.

Figure 9 shows some segmentation results. Since our method is a pure pixel-wise statistical classifier, no filtering is applied. Therefore some segmented hands have holes or background dots. Figure 9g, 9h show the user wearing gloves. Since the gloves have consistent color, they are easily recognized. Figure 9k is an example where the hand Gaussian mistakenly 'grabs' a nearby background color peak during restricted EM training. Figure 9l shows another case where our method fails. This is because the hand color in this image is not consistent: the image was taken inside a car, the thumb and part of the palm was rendered bluish under the windshield. Therefore a single Gaussian can no longer model the hand color, which leads to the failure. Nonetheless, given the difficulty of the test set, we consider our method to be promising.

We are interested in the upper bound performance of the proposed method. That is the performance when we have 'perfect' hand and background color models. Since the test data are also manually segmented, we are able to build a 'perfect' hand color model for each test image by normalizing its hand color histogram:

and similarly a 'perfect' background color model. Then we use these 'perfect' models to classify pixels in the same image. The performance is considered the upper bound of our method, because the models obtained by (6) and (7) are approximations to these 'perfect' models.



We compare the Bayesian method with a simple skin color segmentation method that is frequently (but often implicitly) used. The simple method builds a hand color model and classifies a pixel as a hand pixel if

for some threshold , otherwise it classifies it as a background pixel.

We are interested in the upper bound performance of this method. As in Section 4.2 we build a 'perfect' hand color model for each image, and classify the pixels in the same image with the simple method. We repeat the experiment with different thresholds . Figure 10a shows the performance curve with different . The lowest error rate is 18.3% with =0.015. Figures 10b and 10c show how false negative and false positive rates change with respect to in the method. (the color distributions are mapped to one dimension). Since the two distributions are intrinsically overlapping, a small will generate less false negative but more false positive, and vise versa.

Table 1 summarizes the performances of different methods. Obviously the Bayes decision method is better than the simple threshold method.

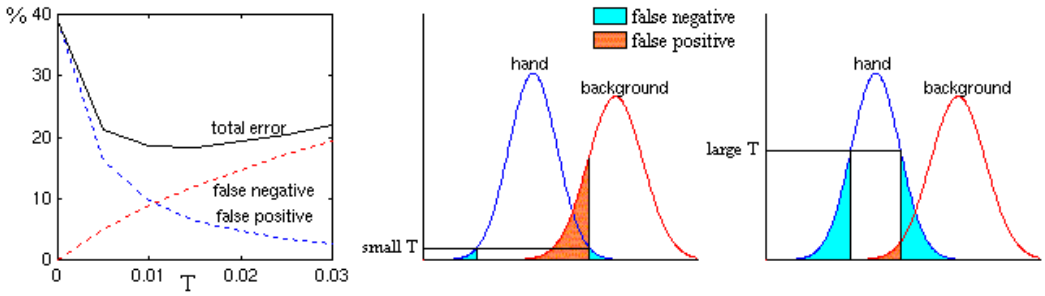
	False Neg.	False Pos.	Error
Bayes Method (Actual)	7.5%	4.0%	11.6%
Bayes Method (Upper Bound)	3.1%	3.6%	6.7%
Simple Method (Upper Bound)	6.4%	11.9%	18.3%

5. Conclusions

We proposed a new way of color segmentation for hand recognition in a wearable environment. The method builds statistical hand and background color models for each image using GMM and the restricted EM algorithm, and classifies pixels with Bayes decision criterion. The performance of the proposed method is promising.

The success of this method relies on the assumption that hand color in a given image is consistent, and hence can be modeled by a Gaussian distribution. Another important prerequisite is that there need to be a few positions where hand tends to occur with high probability, so that the average hand color in a given image can be estimated reliably. The wearable computer application mentioned in Section 2 satisfies these requirements.

Many things can be done for further improvement. For example, some conventional image processing methods such as filtering and region growing will definitely help. More over, currently each pixel is processed individually, whereas it might be beneficial to consider interactions between pixels. In addition, we only considered color information. As the upper bound performance reveals, there is a limit on how



well we can do with color. Adding different information, such as shape, would be helpful. We are investigating some of them, and are applying the proposed method to create a gesture based wearable computer interface.

6. Acknowledgement

The authors would like to thank Larry Wasserman, Roni Rosenfeld, Hua Yu, Pan Yue, Ke Yang, Iain Matthews, William Kunz and all members in the Interactive Systems Labs for their inspiring suggestions and help. This research is partially supported by the Defense Advanced Research Projects Agency under contract number DAAD17-99-C-0061. Xiaojin Zhu's research is supported in part by the National Science Fundation under grant SBR-9720374.

References

[1] V. Pavlovic, R. Sharma, T. S. Huang. *Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review.* IEEE PAMI, Vol 19, No.7, pp. 677-695, 1997

[2] T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R. Picard, A. Pentland. *Augmented Reality Through Wearable Computing.* Presence Vol. 6, No. 4, 1997.

[3] S. Ahmad. *A Usable Real-Time 3D Hand Tracker.* Proceedings of the Twenty-Eighth Asilomar Conference on Signals, vol.2, pp. 1257-1261, 1995.

[4] K. Imagawa, S. Lu, S. Igi. *Color-Based Hands Tracking System for Sign Language Recognition.* Proc. 3rd Int'l Conf. On Automatic Face and Gesture Recognition, pp 462-467, 1998

[5] R. Kjeldsen, J. Kender. *Finding skin in color images.* Proc. 2nd Int'l Conf. On Automatic Face and Gesture Recognition, pp. 312-317, 1996.

[6] Y. Raja, S. J. McKenna, S. G. Gong. *Tracking and Segmenting People in Varying Lighting Conditions using Colour.* Proc. 3rd Int'l Conf. On Automatic Face and Gesture Recognition, pp 228-233, 1998

[7] D. Saxe, R. Foulds. *Towards Robust Skin Identification in Video Images.* Proc. 2nd Int'l Conf. On Automatic Face and Gesture Recogntion, pp. 379-384, 1996.

[8] T. Starner, J. Weaver, A. Pentland. *Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video.* IEEE PAMI, vol.20, no.12, pp. 1371-1375, 1998.

[9] J. C. Terrillon, M. David, S. Akamatsu. *Automatic Detection of Human Faces in Natural Scene Images by Use of a Skin Color Model and of Invariant Moments.* Proc. 3rd Int'l Conf. On Automatic Face and Gesture Recognition, pp.112-117, 1998

[10] M. H. Yang, N. Ahuja. *Extraction and Classification of Visual Motion Patterns for Hand Gesture Recognition.* Proc. IEEE CVPR, pp. 892-897, 1998

[11] J. D. Foley, A. van Dam, S. K. Feiner, J. F. Hughes. *Computer Graphics: Principles and Practice.* 2nd ed. p. 590. Addison-Wesley, Mass., 1993

[12] R. O. Duda, P. E. Hart. *Pattern classification and scene analysis.* Wiley, NY, 1973

[13] A. P. Dempster, N. M. Laird, D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, 39 No.B:1-38, 1977

[14] T. Yamazaki. *Introduction of EM algorithm into color image segmentation.* Proc. ICIPS'98, pp. 368-371, Aug. 1998

[15] L. Wasserman. *personal communications.* 1999