Two Hand Tracking using Colour Statistical Model with the K-means Embedded Particle Filter for Hand Gesture Recognition

Surachai Ongkittikul Stewart Worrall Ahmet Kondoz
School of Electronics and Physical Sciences, University of Surrey, Guildford, Surrey GU2 7XH, UK
s.ongkittikul@surrey.ac.uk s.worrall@surrey.ac.uk a.kondoz@surrey.ac.uk

Abstract

Particle filtering is an efficient and success technique for tracking 2D and 3D through an image. We presents the enhanced tracking of two hands based on a statistical model using only a skin colour feature with particle filtering for gesture recognition. Our framework employs two particle filters to locate two hands individually with the likelihood of the skin pixel-wise classification in the window search. Skin classifier decision was trained from a set of skin samples in YCrCb space using an elliptical model. The tracking scheme employs the reliability measurement derived from the particle distribution which is used to adaptively weight the colour classification. K-means algorithm is used to discriminate the split and merge between left and right hand. Experiments with a set of videos including the movement of two hands in cluttered backgrounds show that adaptive use of our scheme provides improvement compared to use with other techniques such as mean-shift tracking.

Index Terms: Hand tracking, Particle filter, Mean-shift, K-means, HCI.

1. Introduction

Computer becomes more pervasive in modern life. Users require a convenient, natural, unlimited boundary and efficient interaction between human and machine known as human-computer interaction (HCI). Vision-based gesture recognition is an active area for improving HCI [1] such as virtual reality (VR) rather than equipment usage such as mouse and keyboard. Obviously, the human hand is articulated, and it is difficult to capture its non-rigid motion due to its continuously changing shape. Some methods use a data glove to capture hand movement, but they are uncomfortable for the users. Vision-based methods can be used without any restrictions upon the user. Hand tracking is a vital segment for gesture recognition, in which the signer's or user's hands must be detected and localized in image frames.

In general, hand tracking involves the capture of hand motion in 2D image sequences which is performed in the 3D domain. Then tracking can be carried out in the 3D domain or in the 2D image plane. For 2D domain, hands are represented by their geometric features such as shape contour [2]. By using geometric features [3], Isard and Blake attempted to use B-spine curves to model hand contours. Finger tips are another effective geometric feature for hand tracking, for example, multi finger tracking is employed for gesture recognition by [4].

Two hand tracking has been employed in multiple tracking problems which cope with deformable and, in some cases; indistinguishable targets. Multiple cues such as colour, texture and edges are used as a group of features for tracking objects using a particle filter [5]. Tracking multi faces has been demonstrated in [6] based on probability hypothesis density.

The proposed work is part of an ongoing project on a cash machine simulator [7] and expected to design and implement a realistic interface for tracking two hands via vision-based access to the cash machine using gesture recognition. To achieve efficient and robust hand tracking, there are many problems to be solved such as occlusion, abrupt motion, and clutter background. In most of the previous developed tracking techniques, there are two main approaches: top-down and bottom-up. The top-down approach employs an object (model) hypothesis and tries to verify it using the image whereas the image is segmented into objects which are then used for tracking in the bottom-up approach. This paper follows the bottom-up approach. However, to increase the efficiency of hand segmentation, weight function of a pixel classification is included which has been adapted from color distribution [8].

We propose two hand trackers which can undergo non-rigid deformations, rotations or rapidly movement, ambiguity between two hands and interference with clustering such as signer's face. Considering simplicity and practical feasibility, we perform the hand tracking based on the 2D image plane and use the least possible knowledge to reduce the calculation cost and assumptions. Previously, many methods have been developed. Most of them are designed for a specific problem. In this paper, we apply the tracker for gesture recognition in clutter backgrounds or non controlled environments to the cash machine simulator. The presence of background clutter, abrupt motion and varying illumination means that hand tracking is a typical non-linear and non-Gaussian problem. To overcome this, a particle filter [9] was employed.

Particle filtering is also known as sequential Monte Carlo filtering, is the most popular approach and recursively constructs the posterior probability distribution function of the state space using Monte Carlo integration. It has been developed and applied to the hand tracking problem and is also know as CONDENSATION [3].

The structure of this paper is as follows. The proposed skin classifier approach is described in section 2. In section 3, the particle filtering scheme is presented and K-means method for discriminating the right and left hands is described. In section 5, results and comparison with others algorithms is illustrated. Finally in Section 6 we draw the conclusions.

2. Skin Classification with CrCb colour space

The hand tracker employs various techniques to discriminate between object and background. Skin colour is a popular feature that has been used to track hands. Skin classification is involved in this task. It is important to choose an appropriate colour space for skin-colour classification. Colour spaces used in the past have included the YCbCr [10], HSV, RGB, etc. Note that in the RGB colour space, the luminance component and the chrominance components are not decoupled. Also, the feature space is 3D for RGB as opposed to 2D for chrominance skin-colour classification. We considered the YCbCr colour space in our work since it is effective in modelling human skin-colour. By considering the chrominance components only, the feature space is reduced from 3D to 2D, thus reducing the computational complexity of the classification algorithm. In practice, there are many difficulties in perfecting the skin pixel classifier due to the high number of different skin types, changes in ambient light and object movement. The main problem of the classifier is how to model the skin colour and build the decision rules and how to discriminate between skin and non-skin pixels. The simplest scheme is through an explicitly defined skin region. Obviously, two popular groups of skin model are non-parametric and parametric distribution model. There are advantages and disadvantages to these models. Non-parametric models are quickly trained and are independent of the shape of skin distribution but they store the whole training data. Parametric models, on the other hand, do not necessarily carry the whole training data and they are easy to adapt. We selected parametric modelling for these reasons.

Some well-known parametric models are the single Gaussian model [11], the mixture of Gaussian model [12], and histograms. All of them have various drawbacks. They require a lot of training, are not feasible for use in changing illumination conditions and they require costly calculations. In our approach, we employed an elliptical boundary model adapted from [13] for skin modelling. This model can be easily constructed from the training data set, and hence its performance is better than the other concepts mentioned above in term of speed and simple in training and evaluation as the single Gaussian model.

From the elliptical boundary model, the ellipse of the skin colour distribution can then be found and scaled to the CbCr space from training process with a set of skin sample class. A set of elliptical parameters (a, b, ξ_0, η_0 and θ) are found and shown in Figure 1. By comparing the E-distance on the semi-major and semi-minor axes against the skin ellipse boundary, any pixel in CrCb space is classified as a skin pixel if that pixel locates in the skin ellipse boundary (otherwise it is classified as non skin).

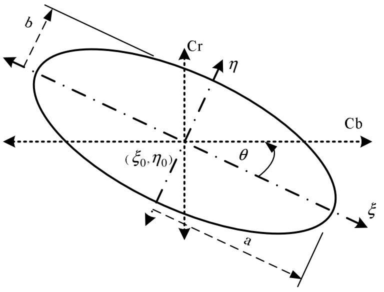


Figure 1. Elliptical parameters

The final decision of the skin classifier is made using a spatial diffusion scheme. In the decision rule, a given pixel will be a skin-pixel if and only if its Euclidean distance, calculated in the CbCr space, with a direct diffusion-neighbour that already belongs to the skin class, is smaller than a threshold. The seed of the diffusion process measures the similarity between neighbour pixels (8-connectivities) as the Euclidean distance.

4. K-means for merge and split hands

In our approach, we start that two hands have to be tracked. Tracker can deal with ambiguity between both hands. This problem will be introduced as merge and split condition between two hands. How trackers cope when two hands move to close together without hesitate to discriminate which one is left or right hands. To break down this problem, we employ the data grouping scheme such as K-means clustering techniques. However, K-means will be used when two hands approach within the threshold range.

K-means clustering [14] partitions a set of data into k sets. We use K-means techniques to consider the set of data which contains the sampling weight in arbitrary position in image. The solution is then set k (equal two for our approach) centers; each of them is located at the centroid of the data which is the closet centre. Define a *d*-dimensional set of *n* data points $X = \{x_1, ..., x_n\}$ as the data to be clustered and define a *d*-dimensional set of *k* clusters $C = \{c_1, ..., c_k\}$. The proportion of data point x_i is defined a membership function $m(c_j | x_i)$ that belongs to center c_i . The membership function consists of two types: hard member function $m(c_j | x_i) \in \{0,1\}$ and soft member function $0 \le m(c_j | x_i) \le 1$. A weight function $w(x_i)$ represents the influence data point x_i to the next iteration of the centre parameters. The general k-means model of iteration is:

1. Initialize the k-means algorithms with the number of clustering.

2. For each data x_i , compute its membership $m(c_j | x_i)$ in each center c_j and its weight $w(x_i)$.

3. For each center c_j , re-calculate its location from all data x_i according to their membership and weights:

$$c_j = \frac{\sum_{i=1}^{n} m(c_j | x_i) w(x_i) x_i}{\sum_{i=1}^{n} m(c_j | x_i) w(x_i)} \tag{1}$$

4. Repeat step2 and 3, stop iteration when conversed.

The generic version of the partitioned clustering is the objective function that the k-means algorithm optimizes is:

$$KM(X, C) = \sum_{i=1}^{n} \min_{j \in \{1...k\}} \left\| x_i - c_j \right\|^2 \tag{2}$$

Objective function (2) gives the minimization of the squared distance each center and its assigned data points. The membership and weight function of k-means algorithm are:

$$m_{KM}(c_l | x_i) = \begin{cases} 1; & if \;\; l = \arg\min_j \left\| x_i - c_j \right\|^2 \\ 0; & otherwise \end{cases} \tag{3}$$

$$w_{KM} = 1 \tag{4}$$

We define K-means is a hard membership function, and a constant weight function in our approach. Obviously, the numbers of clusters consist of two groups in case left and right hands move near together or any hand move near face. However, the numbers of clusters can change to be three (k) depending on the object position for example, both hands move near face.

5. Particle Filtering

Particle filtering is a Bayesian sequential importance sampling technique, which recursively approximates the posterior distribution $p(X_t | Z_t)$ by using a finite set of weight samples. It consists of two stages: prediction and update. The tracking state is described by X_t while the vector Z_t denotes all observations $\{z_1, ..., z_t\}$ up to time *t*.

The filter in Eq.(5) is for approximating the probability distribution by a weight sample set:

$$S = \{(s^{(n)}, \pi^{(n)} | n = 1, ..., N)\} \tag{5}$$

Each sample of the distribution represents a window search area (M pixels) and is given as $s = \{x, y\}$ where *x*, *y* are the center location of the search window. The sample set is propagated through the dynamic model.

$$s_t = As_{t-1} + w_{t-1} \tag{6}$$

Where *A* is identity matrix (assuming initially an random walk) and w_{t-1} is a multivariate Gaussian random variable. Skin distributions are used as target models which are computed in the search window.

$$P_y = f \sum_{i=1}^{M} k\left(\frac{\left\| y - x_i \right\|}{d} \right) C_i \qquad ;\text{when } k(r) = \begin{cases} 1 - r^2 \\ 0 \end{cases} \tag{7}$$

where *M* is the pixel number in the search window, C_i is skin classification and $d = \sqrt{H_x^2 + H_y^2}$ is the search window size, and *f* is a normalization factor. The likelihood function of each sample is defined as.

$$\pi^{(i)} = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(1-P_i)^2}{2\sigma^2}} \tag{8}$$

The likelihood of the particle $s^{(i)}$ is specified by a Gaussian with variance σ which is an empirical constant selected by observing the tracking performance in the test videos.

A common problem with the particle filter is the degeneracy which occurs when, after some iteration, the weights of most particles are very low and disappear. In order to eliminate this problem and to focus on high weight particles, the re-sampling scheme in [15] is employed.

We employ the hand tracking according to locate the hand positions for gesture recognition. Gesture commands are designed using left and right hands with contained the rapid movement and deformable hand shape. As early mention, colour information is used to solve the ambiguity problem between two hands as well as signer face. The particle filter frame work and embedded the K-means clustering are used in this paper. With problem such as merge and split hands (two object trackers) problem when they are closed together by grouping the weight samples. Moreover, K-means will be required if any hand moves close or in font of the face less than the distance threshold (T). The programming details for one iteration step are the embedded K-means particle filter given in Figure 2.

Given the sample set. $S_{t-1} = \left\{ \left(s_{t-1}^{(i)}, \pi_{t-1}^{(i)} \right) | i = 1,..., N \right\}$

and, perform the following steps:

1. Propagating each sample from set S_{t-1} by dynamic model: $s_t^{(n)} \sim p\left(X_t \middle| X_{t-1} = s_{t-1}^n \right)$ to give S_t .

2. Observe the skin distribution:

$$P_{s_t^{(n)}} = f \sum_{i=1}^{M} k \left(\frac{\left\| s_t^{(n)} - x_i \right\|}{d} \right) C_i \qquad ;\text{when } k(r) = \begin{cases} 1 - r^2 \\ 0 \end{cases}$$

3. Selective re-sampling: if any weight sample less than a weight threshold.

4. K-means clustering applies to group the data set if $\left\| s_{t-1}^R - s_{t-1}^L \right\| \le T$; T = Distance threshold between two hands.

5. Estimate the mean state of the set S_t

$$E[S_t] = \sum_{n=1}^{N} \pi_t^{(i)} s_t^{(i)}$$

Figure 2. An iteration step of the particle filter

5. Experimental Results

In hand tracking experiments, we begin by training our skin colour model and optimize the elliptical parameters by using probability of false alarm and miss [16]. The performance of the hand tracker is demonstrated by comparing with Mean-shift technique.

5.1. Skin Colour Modelling

The skin classifier has been trained by the set of skin and non-skin samples. We selected a set of skin colour samples (altogether about 200,000 pixels) from a total of 10 signers. Those signers consist of the variety of nationalities. To construct the elliptical model, all the skin pixels are set within an elliptical boundary. In Figure 3, it is shown that the elliptical model covers all skin pixels but includes some blank areas as well. To justify the pixel classification, we use probability of false alarm and miss. The results had shown that scale of the size ellipse against P_{error}. Where priori probability of skin class $P(\omega_S)$ = 0.15 and non-skin class $P(\omega_{\bar{S}})$ = 0.85. The minimum probability of error is corresponds to scale = 0.75.

5.2. Tracking results

To illustrate the performance of our approach, we had done experiment by applying our methods to the videos. Firstly, we define two-hand gesture commands using for cash machine simulator [7] followings about 10 gestures, e.g., print balance, see balance, amount money, begin, withdraw. For example, print balance and see balance commands:

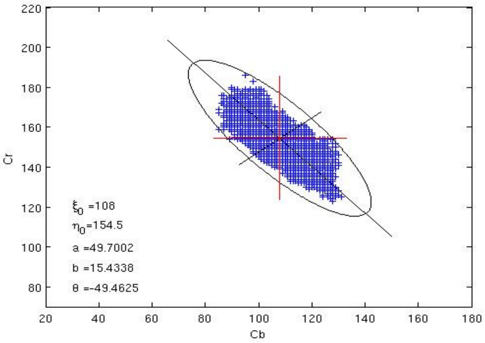
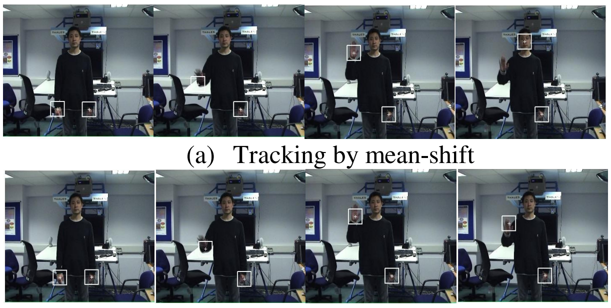


Figure 3. Skin distribution with elliptical model

Print balance: Both hands are held in front of the body. The right hand moves towards to the left hand so that the right hand palm touches the left hand palm; the right hand turn away from the signer and to the right. Both hands are back in front of the body, move up and down alternatively several times and finish in the neutral space.

See balance: The right hand with point two fingers is held with the palm facing away from the signer. The right hand moves up and held in front of the face and then moves back to beginning and both hands are back in front of the body, move up and down alternatively several times and finish in the neutral space.

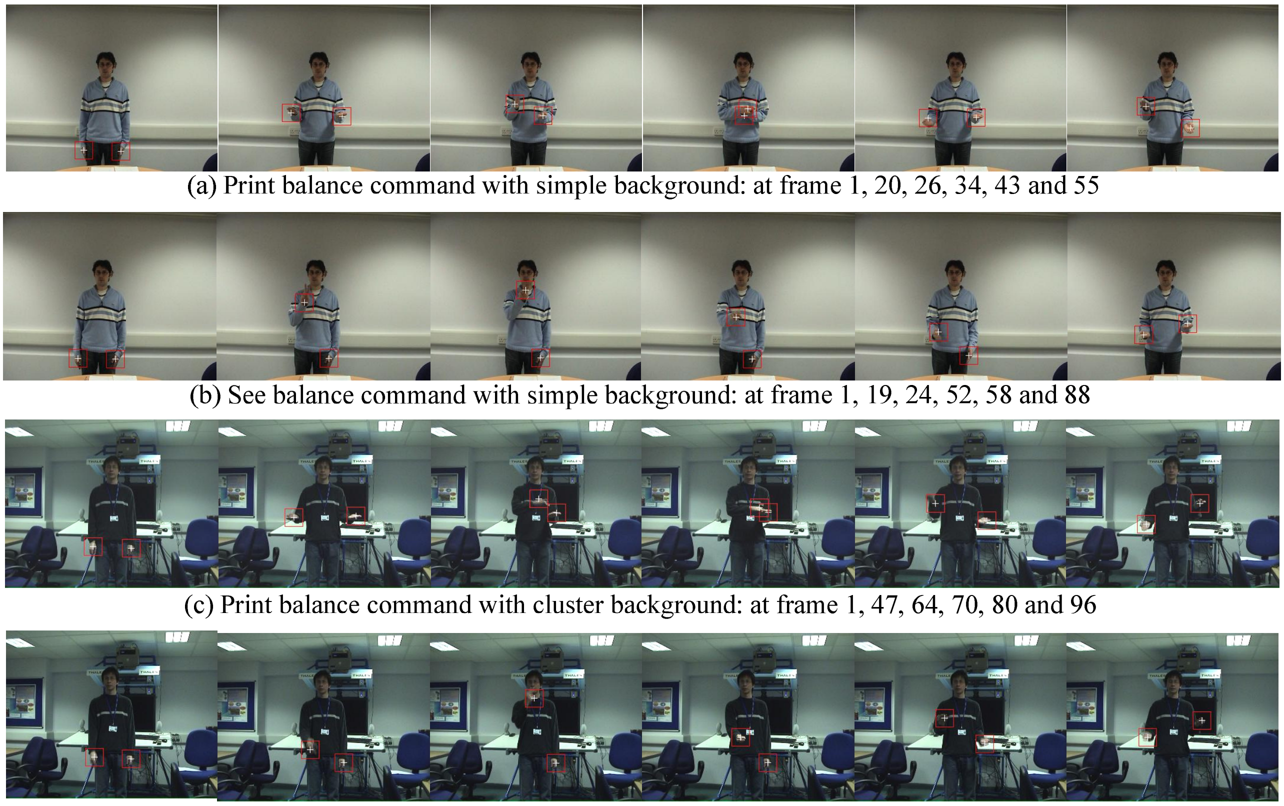
We performed the hand tracking experiments on more than 45 videos of hand gestures and each video contains around 60 to 130 frames. These video sequences are captured at 20 frames per second and the resolution is 320x288 for each frame. The initial positions of both hands are located at the first frame of video. Each tracked window of particle filter is 30x30 pixels and the number of particles is 40.



(a) Tracking by mean-shift

(c) Tracking by enhanced particle filter

Figure 4. Tracking results compare between with our approach with Mean-shift with ambiguity of hand and signer face



(a) Print balance command with simple background: at frame 1, 20, 26, 34, 43 and 55

(b) See balance command with simple background: at frame 1, 19, 24, 52, 58 and 88

(c) Print balance command with cluster background: at frame 1, 47, 64, 70, 80 and 96

(d) See balance command with cluster background: at frame 1, 9, 20, 48, 57 and 64

Figure 5. Two hands tracking with k-means embedded particle filter with simple and cluster background.

The simple and cluttered background was also used to set up these video sequences. First experiments shown in Figure 4, we apply our approach to the video sequences containing hand with abrupt motion and moving close to face compare with mean-shift tracker. The results are shown for captured frames at frame instants 1, 15, 25 and 49. Mean-shift tracker looses the hand due to the confusion face area at frame 25. Our scheme can deal with the rapid motion and ambiguity of skin face nearby hand till the end video.

Next, the sequence of pictures in Figure 5 shows the tracking sequences of two hands which contained the rapid motion and ambiguity between both hands and signer face with the 'Print balance' and 'See balance' commands, respectively. Figure 5 (a-b), those videos captured over white background and Figure 5 (c-d) captured over cluster background. The tracker can track followed both hands with robustness. However, 83% of video worked well and 17% lost track. The significant reason of lost tracking is too quick moving of hands.

We observe from the frame that tracking is beginning to be lost from the target hand. The hand shape in that image frame is very blur and un-clear as well as the colour change by mixing with the background around that hand. Then, the skin colour classification is the important state to increase the potential of tracker.

6. Conclusion and future work

In this paper, we propose an efficient and robust tracing of two hands tracker with embedded k-means particle filter for cash machine simulator by designed the particular hand gesture for this task. Two particle filters track each hand and employ K-means algorithm to discriminate the sample weight in that merge and split conditions between two hands. Each particle filter takes the sample set to estimate the likelihood function based on colour classification. Elliptical boundary model used to define the skin clolour in CbCr space. Skin training is required before tracking. In tracking experiment, we test our approach with videos which contain the simple and cluster background. The experiment results show that the performance of our tracker is work about 83% of all videos. Reason of lost track come from the hand moved too quickly. The shape and colour are distortion as blurred. To enhance to decreases the lost track chance, by including more observation models with skin colour features or change the colour classification to be adaptive the tracking performance will be improved greatly.

However, this can be inefficient or infeasible when a target exhibits abrupt motion or when there is an ambiguity in the target. In this case, cues in the observation model from the hand tracking are essential.

7. Acknowledgements

The work presented was developed within VISNET II, a European Network of Excellence (http://visnet-noe.org), funded under the European Commission IST FP6 program.

8. References

[1] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review," *IEEE Trans. on PAMI,* vol. 19, pp. 677-695, 1997.

[2] G. McAllister, S. J. McKenna, and I. W. Ricketts, "Hand Tracking for Behaviour Understanding," *Image and Vision Computing,* vol. 20, pp. 827-840, 2002.

[3] M. Isard and A. Blake, "CONDENSATION conditional density propagation for visual tracking," *International Journal of Computer Vision,* vol. 1, pp. 5-28, 1998.

[4] K. Oka, Y. Sato, and H. Koike, "Real-Time Fingertip Tracking and Gesture Recognition," *IEEE Computer Graphics and Applications,* vol. 22, pp. 64-71, 2002.

[5] M. Jaward, L. Mihaylova, N. Canagarajah, and D. Bull, "Multiple Object Tracking Using Particle Filters," in *Aerospace Conference*, 2006.

[6] E. Maggio, E. Piccardo, C. Regazzoni, and A. Cavallaro, "Particle PHD Filtering for Multi-Target Visual Tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, 2007.

[7] K. Ignasiak, M. Morgos, and S. Ongkittikul, "Architecture of Information System for Intelligent Cash Machine," *International Conference Signal Processing and Multimedia Applications (SIGMAP 2007),* pp. 28-31, 2007.

[8] K. Nummiaro, E. Koller-Meier, and L. V. Gool, "An Adaptive Color-Based Particle Filter," *Image and Vision Computing,* vol. 21, pp. 99-110, Jan 2003.

[9] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A Tutorial on Particle Filters for On-line Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Transactions on Signal Processing,* vol. 50, pp. 174-188, 2001.

[10] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Transactions Circuits and Systems Video Technology,* vol. 9, pp. 551-564, Jun 1999.

[11] L. Fan and K. K. Sung, "Face detection and pose alignment using colour, shape and textureinformation," in *Third IEEE International Workshop on Visual Surveillance*, pp. 19-25, 2000.

[12] Y. Raja, S. J. McKenna, and S. Gong, "Tracking and Segmenting People in Varying Lighting Conditions using Colour " in *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 228-233, 1998.

[13] J. Y. Lee and S. I. Yoo, "An Elliptical Boundary Model for Skin Color," in *Proc. of the Int. Conf. on Imaging Science, Systems, and Technology.* , 2002.

[14] G. Hamerly and C. Elkan, "Alternatives to the k-means algorithm that find better clusterings," in *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 600-607, 2002.

[15] M. Isard and A. Blake, "ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework," in *Proceedings of the 5th European Conference on Computer Vision*, pp. 893-908, 1998.

[16] 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,* vol. 14, pp. 1086- 1097, 2004.