# Two Hand Dynamic Gesture Recognition Using Random Sampling Techniques

Sarita M Bopalkar
Department of Electronics & Telecommunication
Smt Indira Gandhi College of Engineering
Navi Mumbai, India
saritabopalkar@gmail.com

## ABSTRACT

This work develops a framework for recognition of two hand dynamic gestures, using condensation algorithm. The work is broadly divided into three parts. First part of this work deals with skin color identification using color segmentation using' Gaussian Mixture Model'. In the Second part hand motions are modeled as trajectories of some estimated parameters over time. During training, one template trajectory and one prototype feature vector are generated for every gesture class. Features used in this work include some static and dynamic motion trajectory features. The third part of this algorithm is to recognize the input gesture. Random sampling techniques with 'Condensation Algorithm' are used for incrementally comparing and matching the trajectory models to the input gesture. The recognition framework is demonstrated with examples which include two hands dynamic gestures. The gestures considered for the development of algorithm are generic in nature and can be applied to various applications where information is not conveyed through speech, like space gestures, under water gestures, airplane parking signs, sign language gestures, etc.

## General Terms

Image Processing & Computer Vision, Segmentation, Pixel Classification, Human Factors.

## Keywords

Human Machine Interaction; Gaussian Mixture Model; Motion Trajectory; Condensation; dynamic hand gesture

## 1. INTRODUCTION

Gesture recognition is a topic in computer science and language technology with the goal of interpreting human gestures via mathematical algorithms. Gestures can originate from any bodily motion or state but commonly originate from the face or hands.Gesture recognition enables humans to interface with the machine (HMI) and interact naturally without anymechanical devices. Gesture recognition can be conducted with techniques from computer vision and image processing. Vision based handgesture recognition involves visual analysis of hand shape, position and or movement. One notable advantage of vision based system is that it is easy to deploy as only camera is sufficient to capture the input from the user and user does not need to ware special hardware likehand gloves orattach markers [1]. Thus computer vision has more natural interface with machines.There are different tools existing for gesture recognition. Most of the problems in gesture recognition have been addressed based on statistical modeling tools, such as PCA, HMMs [2], Kalman filtering [3], more advanced particle filtering [4] and condensationalgorithms [5].The Condensation algorithm (Conditional Density Propagation over time) found more effective as it makes use of random sampling in order to model arbitrarily complex probability density functions.

In this work, motion trajectories of hands are used to recognize gesture. It attempts to use particle filter to classify the gesture. The designed gesture recognition algorithm uses skin modeling for skin color segmentation, using Gaussian Mixture Model to identify hands. Further the hand motions are modeled as trajectories with estimated parameters over time. Condensation algorithm is applied for gesture recognition. This paper also briefs about the experimental results and performance evaluation of overall algorithm.

## 2. SKIN COLOR SEGMENTATION

First step towards this algorithm is to capture the image sequence for particular gesture. Then localize hand and face in the image and segment from the background before recognition. Skin-color information is considered for segmentation because it is a very effective tool for identifying/classifying hand and facial areas as skin color provides computationally effective yet, robust information against rotations, scaling and partial occlusions. However there are some difficulties in robustly detecting the skin color. The ambient of the light and shadows can affect the appearance of skin tone color. Moreover different camera, ethnicity can affect skin color distribution [6].

The main part in skin color segmentation is to choose the suitable color space. Although skin colors of different individuals appear to vary over a wide range, they differ much less in color than in brightness.Luminance can be removed from the color representation in thechromatic color space. This work uses HSV chromatic color space; it reduces the effect of uneven illumination [7]. The transformation of RGB to HSV is invariant to high intensity at white lights, ambient light and surface orientations relative to the light source and hence, can form a very good choice for skin detection methods. It is also compatible with the human color perception. The polar coordinate system of Hue-Saturation spaces, resulting cyclic nature of the color space makes it inconvenient for parametric skin color models that need tight cluster of skin colors for best performance. A different representation of Hue-Saturation using Cartesian coordinates can be used [8] given by equation (1):

$$X = S \cos H \, , Y = S \sin H \qquad (1)$$

## 2.1 Skin Modeling

Skin modeling is used to model the distribution of skin and non-skin color pixels. Two different approaches are used for skin modeling. Nonparametric methods include normalized lookup table and bayes classifier are histogram based methods requires large storage space and not possible to generalize the training data.Threfore parametric method which include Gaussian model has more compact and has ability to generalize the training data.

Here Mixture of Gaussian model is used to model skin color given by equation (2).

$$P(c,\theta) = \sum_{i=1}^{k} \propto_i \frac{1}{\sqrt{(2\pi)^d|\sum_i|}} \times exp\{-\frac{1}{2}(c - \mu_i)^T \sum_i^{-1}(c-\mu_i) \qquad (2)$$

Where k is the number of Gaussians in the mixture model, the parameter set $\theta = \{\propto_i, \mu_i, \sum_i\}_{i=1}^{k}$ is such that $\sum_{i=1}^{k} \propto_i = 1,\ \propto_i > 0,\ \mu_i \in Rd$ and $\sum_i$ is a $d \times d$ positive definite matrix. The parameters of the Gaussian mixture model are estimated using the EM (Expectation-Maximization) algorithm [9]. The EM algorithm is an iterative method to obtain the maximum likelihood estimation of the parameter set by equations (3-5):

$$\propto_i^{new} = \frac{1}{N}\sum_{j=1}^{N} P(i|c_j,\theta_i) \qquad (3)$$

$$\mu_i^{new} = \frac{\sum_{j=1}^{N} c_j P(i|c_j,\theta_i)}{\sum_{j=1}^{N} P(i|c_j,\theta_i)} \qquad (4)$$

$$\sum_i^{new} = \frac{\sum_{j=1}^{N} P(i|c_j,\theta_i)(c_j-\mu_i^{new})(c_j-\mu_i^{new})^T}{\sum_{j=1}^{N} P(i|c_j,\theta_i)} \qquad (5)$$

The first step in applying the EM algorithm is the initialization of the mixture parameters. This is done by initializing the algorithm multiple times with random initial parameters, measuring after each runthe likelihood function. Using equations (3-5), the expectation step andmaximization stepcan be performed simultaneously and the algorithm proceeds byusing the newly derived parameter values as the guess for the next iteration. Using the skin-color model, a likelihood image of the original color image can be computed. The likelihood image is a gray scale image where the gray value of each pixel shows the probability of the pixel to belong to the skin. Examples in fig1 demonstratethe robustness of GMM towards varying illumination condition and cluttered background, towardsage of person and different types of cameras like point and shoot camera or Single Lens Reflex (SLR) camera.

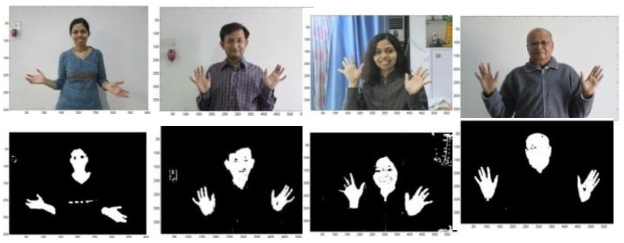


**Fig 1: Skin color segmentation by GMM under different condition.**

The next step is to locate palm of two hands in the image. This is done by region growing algorithm which scans the image from left to right and top to bottom to search two skin blobs for hands.In this work, two hand dynamic gestures that are characterized by total movement of the hands in space are considered. For these types of gestures, hand movement is defined by the motion of the palm in space.Centroids of the hands are considered as a reference point because even if there is some local motion in the fingers, the centroid will be approximately at the center of the palm and hence the coordinate (x,y) of the centroid in a frame will best represent the position of the hand in that frame. So centroid of palm in each frame is recorded.

## 3. Hand Tracking and motion trajectory

Model based method is used here for hand tracking. Centroids of the hands are considered as a reference point. For hands tracking, it is necessary to decide blobs in each frame of given video sequence are whether belong to left or right hand of user. For that the centroids of two blobs in each frame has recorded. And then the first frame in the sequence is examined and centroid farthest to the left and farthest to the right has determined. The one on the left corresponds to the right hand of the user. The one on the right corresponds to the left hand of the user. Then for each successive frame, centroid which is closest to each of the previous left centroid called new left centroid. And the same to determine new right centroid in next frame. Thus centroid of two blobs for left and right hands are separated. The computed centroids are then joined in the temporal sequence to obtain the final estimated gesture trajectory $G$. $G = (x_1, y_1), \ldots, (x_n, y_n)$Where n is the number of frames in gesture sequence. Fig2 shows Left and right hand trajectory for Sample Gesture 1.

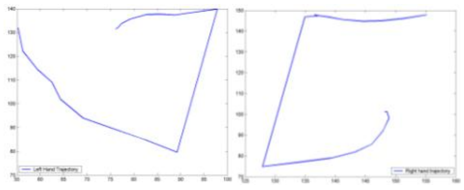


**Fig2: Left and Right Hand trajectory**

## 3.1 Feature Extraction from template trajectories

Trajectory guided gesture recognition requires matching/ comparing of an input gesture trajectory to each of the prototype gesture trajectories contained in the database.Here the gesture is classified as one of the gestures in the database.

Fig 3 and 4 shows example of one sample gesture and its segmented images. Here gestures are characterized by the motion of the hand in space. Hence in order to recognize gestures, features which can capture these gesture characteristics are required.

Here some static and dynamic features are extracted from a given gesture trajectory. Static features relate to the shape of the trajectory while the dynamic features relate to the nature of hand motion during gesturing. The static features considered in this work correspond to the total length of the extracted trajectory. And dynamic features considered here are average horizontal and vertical velocity of both the hands which relate to the motion information. The choice of all these features is based on observations over a large number of gesture samples. It is observed the nature of hand motion when gestures are performed by different persons may be different and the above mentioned features are considered after analyzing their behaviors. To calculate average velocities five traininginstances has considered here. The horizontal and vertical velocities are calculated as shown in equation (6)

$$v_H = x_{t+1} - x_i\,, \quad i = 0,1,2,\ldots\ldots,n-1$$

$$v_V = y_{t+1} - y_i\,, \quad i = 0,1,2,\ldots\ldots,n-1 \qquad (6)$$

Where $v_H$ and $v_V$ are horizontal and vertical velocities, n are total numbers of frames in video sequence of gesture, (x,y) are centroids recorded for both hands.

Then motion trajectory model has created using above feature. The graphs in figure 5 and 6 shows the plot of vertical and horizontal velocities over time for video sequence derived for sample gesture 1 and sample gesture 2 respectively. Specifically, here for each gesture, five training instances are used to calculate the average horizontal and vertical velocities of both hands in that particular frame.

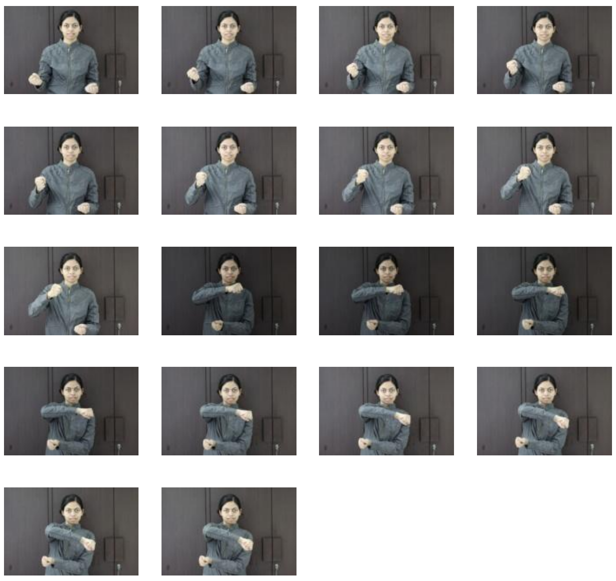
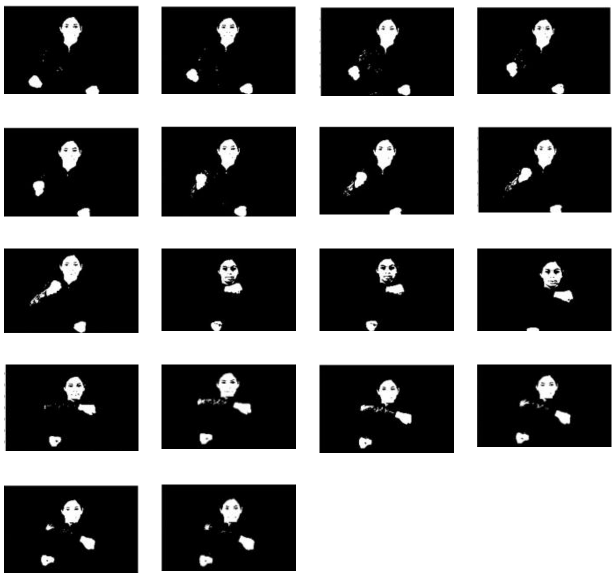


**Fig3: Sample gesture 1**

**Fig4:Segmented Sample gesture 1 (Model-I)**



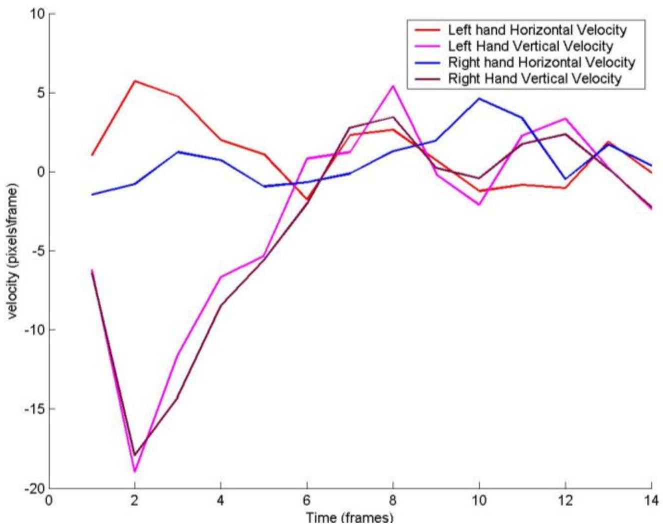

Left hand Horizontal Velocity
Left Hand Vertical Velocity
Right hand Horizontal Velocity
Right Hand Vertical Velocity
velocity (pixels\frame)
Time (frames)
10
5
0
-5
-10
-15
-20
0
2
4
6
8
10
12
14

**Fig5:Motion Trajectory for Sample gesture 1**

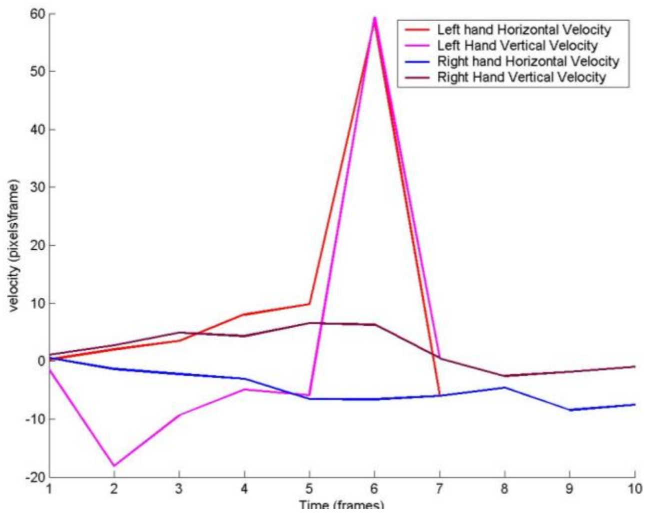


**Fig6: Motion Trajectory for Sample gesture 2**

## 4. Condensation algorithm to classify new video sequences

This section follows a brief description of how the Condensation algorithm is applied to gesture recognition task. The Condensation algorithm (Conditional Density Propagation over time) makes use of random sampling in order to model arbitrarily complex probability density functions [10]. The key strength of particle filter approach is that computational requirement varies linearly with number of particles and its ability to deal with multiple hypotheses.

Here goal is to take a set of M model trajectories $\{m^{(\mu)}, \mu = 1, \dots, M\}$and match them against an N-dimensional input trajectory, given by $z_t = (z_{t,1,\dots,}z_{t,N})$ at time t where $z_t$ are observations at time t. The models are taken to be discretely sampled curves with a phase parameter $\theta \,\epsilon\, [0, \theta_{max}]$ representing the current position in the model. The model values at position $\theta$ are a vector of N values $m_\theta^{(\mu)} = (m_{\theta,1}^{(\mu)}, \dots, m_{\theta,N}^{(\mu)})$ where the stored discrete curve is linearly interpolated at phase$\theta$. Specifically, a state at time t is described as a parameter vector: $s_t = (\mu, \theta^i, \alpha^i, \rho^i)$ where $\mu$ is the integer index of the predictive model, $\theta^i$ indicates the current position in the model, $\alpha^i$ refers to an amplitude scaling factor, $\rho^i$ is a scale factor in the time dimension where $i \in \{l, r\}$, i indicates which (left or right) hand's motion trajectory this $\theta^*, \alpha^*$ or $\rho^*$ refers to. In this work models contain data about the motion trajectory of both the left hand and the right hand; by allowing two sets of parameters, here the motion trajectory of the left hand to be scaled and shifted separately from the motion trajectory of the right hand (so, for example, $\theta^l$ refers to the current position in the model for the left hand's trajectory, while $\theta^r$ refers to the position in the model for the right hand's trajectory). In summary, there are 7 parameters that describe each state.

The algorithm is divided into three basic steps:

Initialization: The sample set is initialized with N samples distributed over possible starting states and each assigned a weight of $\frac{1}{N}$. Specifically, the initial parameters are picked uniformly using according to:

$$\mu \,\epsilon\, [1, \mu_{max}]$$

$$\theta^i = \frac{1-\sqrt{y}}{\sqrt{y}} \quad , \quad \text{where } y \,\epsilon\, [0,1]$$

$$\propto^i \,\epsilon\, [\alpha_{min}, \alpha_{max}]$$

$$\rho^i \epsilon\, [\rho_{min}, \rho_{max}]$$

Prediction: In the prediction step, each parameter of a randomly sampled $s_t$ is used to determine $s_{t+1}$ based on the parameters of that particular $s_t$. Each old state$s_t$, is randomly chosen from the sample set, based on the weight of each sample. That is, the weight of each sample determines the probability of its being chosen. This is done efficiently bycreating a cumulative probability table, choosing a uniform random number on [0, 1], and then using binary search to pull out a sample. The equations used to select new state are shown below:

$$\mu_{t+1} = \mu_t$$

$$\theta_{t+1}^i = \theta_t^i + \rho_t^i + N(\sigma_\theta)$$

$$\propto_{t+1}^i = \propto_t^i + N(\sigma_\propto)$$

$$\rho_{t+1}^i = \rho_t^i + N(\sigma_\rho)$$

where $N(\sigma_*)$ refers to a number chosen randomly according to the normal distribution with standard deviation $\sigma_*$.This adds an element of uncertainty to each prediction, which keeps the sample set diffuse enough to deal with noisy data. In this application it has set as: $\sigma_\theta = \sigma_\alpha = \sigma_\rho = 0.1$.

For a given drawn sample, predictions are generated until all of the parameters are within the accepted range. If, after, a set number of attempts it is still impossible to generate a valid prediction, a new sample is created according to the initialization procedure above.

Update: After the Prediction step above, there exists a new set of N predicted samples which need to be assigned weights. The weight of each sample is measure of its likelihood given the observed data $Z_t = (z_t, z_{t-1}, \dots)$.Here $Z_{t,i} = (z_{t,i}, z_{(t-1),i}, \dots)$ is defined as a sequence of observationsfor the ith coefficient over time; specifically let $Z_{t,1}, Z_{t,2}, Z_{t,3}, Z_{t,4}$be the sequence of observations of the horizontal velocity of the left hand, the vertical velocity of the left hand, the horizontal velocity of the right hand, and the vertical velocity of the right hand respectively.The weight of the state is calculated by equation (7).

$$P(z_t|s_t) = \prod_{i=1}^{4} P(z_{t,i}|s_t) \qquad (7)$$

Where

$$P(Z_{t,i}|s_t) = \frac{1}{\sqrt{2\pi}} exp \frac{-\sum_{j=0}^{w-1}(z_{(t-j),i} - \alpha^* m^{(\mu)}_{(\theta^*-\rho^* j),i})^2}{2(w-1)}$$

Where w is the size of a temporal window that spans back in time (here, w = 10 has taken). Note that $\alpha^*, \theta^*, \rho^*$ refer to theappropriate parameters of the model for the blob and that $\alpha^* m^{(\mu)}_{(\theta^*-\rho^* j),i}$ refers to the value given to the ithcoefficientof the model $\mu$interpolated at time$\theta^* - \rho^* j$ and scaled by $\alpha^*$.

Classification: With this condensation algorithm last step is classifying the video sequence as one of the gesture from database gestures. Since the whole idea in condensation is that the most likely hypothesis will dominate by the end. Here inthis work, this criterion is used to recognize which model was deemed most likely at the end of the video sequence to determine the class of the entire video sequence. For this probability assigned to each model is determined by summing the weights of each sample in the sample set at a given moment whose state refers to the model in question. Experimental results follow the graphs plot for likelihood of each model over time for an instance of each gesture.

In this step a particle filter is used to classify both the test and training sets of video sequences to define a deterministic process of gesture recognition. To reduce the processing time the Condensation algorithm is coded in VC++ and run much faster.

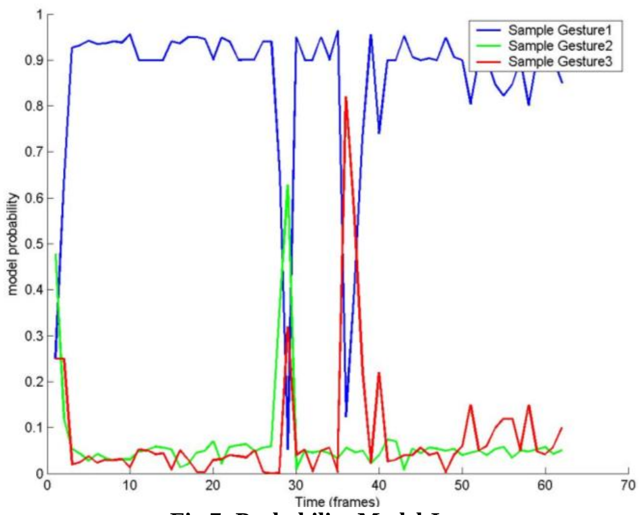


**Fig 7: Probability Model-I**

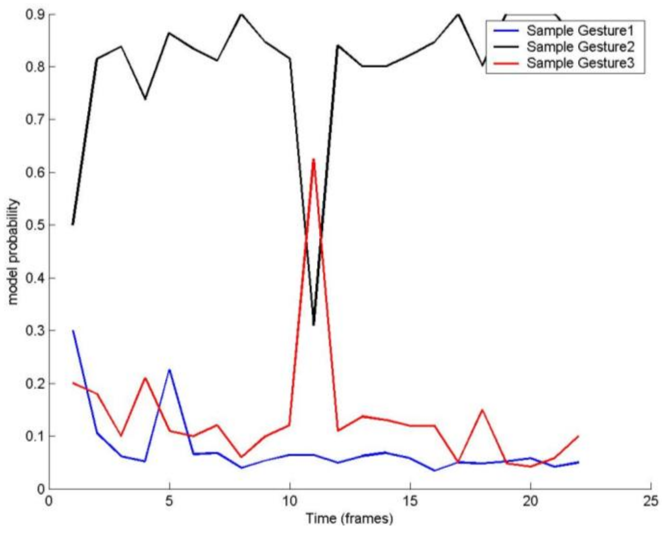


**Fig8: Probability Model-II**

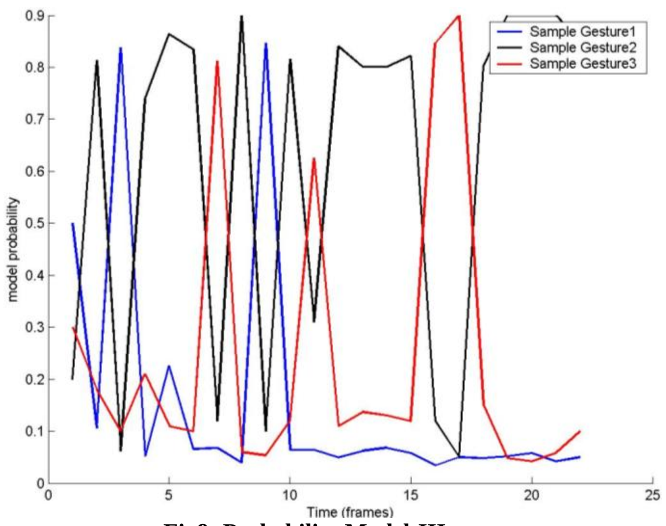


**Fig9: Probability Model-III**

Fig7 shows the probability plot for input test gesture 1. This gesture is same as sample gesture1. Fig8 shows the probability plot for input test gesture 2. This gesture is same as sample gesture 2.Fig9 shows the probability plot for input test gesture 3.This gesture is different from database gestures. These input test gestures are done by different person from training model and also these are captured at different instant of time and under differentillumination condition and background from training model to test the robustness of the algorithm.

Fig7 shows probability plot for the Input Gesture 1 with respect to Sample gesture-1,Sample Gesture-2and sample gesture-3. Heremaximum probability is with Sample Gesture-1 over the time so it is classified as Sample Gesture-1 where as in fig8 Input Gesture is classified as Sample Gesture-2 because maximum probability is with Sample Gesture-2 over the time.Like this other input gesturers has tested which shows the robustness of the algorithm with high recognition rate. Fig9 shows the probability plot for the input test gesture which not from training data. These gesture probabilities show that these gestures don’t match with Sample Gesture-1 or Sample Gesture-2 or Sample Gesture-3. Therefore recognized as a gesture not from training data base.

## 5. Conclusion

In the first part of the work, skin detection using color segmentation has been attempted. While Gaussian Mixture Model has given consistent results for cluttered background as well as different illumination conditions.

A model-based method for tracking hand motion in space is used. Some of the static and dynamic features extracted from the gesture trajectory are co-related to the shape as well as the nature of hand motion during gesturing. These features are selected based on the different observations.

The particle filters are successfully used for the recognition of the gesture. The random sampling used by the Condensation Algorithm is found appropriate in modeling arbitrarily complex probability density functions.

The algorithm is also found robust in identifying gestures made by different persons, at different illumination and with different backgrounds as illustrated in the experimental results.

The future work involves further challenge is to optimize the algorithm, code, as well as the computational load to achieve better real time performance.

## 6. References

[1] H. Brashear, T. Starner, and a. H. J. P. Lukowicz, "Using multiple sensors for mobile sign language recognition," Proceedings of IEEE International Symposium on Wearable Computers,, pp. 45{52,, October 2003.

[2] J. Yamato,and J. Ohya, and K. Ishii, "Recognizing human action in time sequential images using hidden Markov model" Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn., Champaign, IL, pp. 379{385, 1992.

[3] G. Welch and G. Bishop, "An introduction to the Kalman filters" Dept. Computer. Sci., Univ. North Carolina, Chapel Hill, Tech., vol. Rep. TR95041, pp. 677-695, 2000.

[4] C. Kwok,and D. Fox, and M. Meila,"Real-time particle ¯lters" Proc. IEEE, vol. 92, no. 3, pp. 469{484,, Mar. 2004.

[5] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density"," Proc. Eur. Conf. computer. Vis., Cambridge, U.K., pp. 343-356, 1996.

[6] Beetz, M., and B. Radig, and M. Wimmer, "A person and context specific approach for skin color classification"18th International Conference on Pattern Recognition, 2006 (ICPR 2006).Hong Kong, 2006.

[7] M. Soriano and et.al.,"Skin detection in video under changing illumination conditions" 15th International Conference on Pattern Recognition 2000.Barcelona.2000.

[8] N. Oliver,and A. Pentland, and F. Berard,and Lafter, "lips and face real time tracker," CVPR97, 1997.

[9] TERRILLON, J.-C., SHIRAZI, M. N., FUKAMACHI, H., AND AKAMATSU, S. 2000. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In Proc. of the International Conference on Face and Gesture Recognition, 54–61.

[10] A. Black and M. Isard, “Condensation-conditional density propagation for visual tracking," in press, pp. 1358-1366, (1997).