Real Time Hand Detection & Tracking for Dynamic Gesture Recognition

Varsha Dixit
Department of Information Technology, IIIT Allahabad, UP, India
E-mail: varshadixit009@gmail.com

Anupam Agrawal
Department of Information Technology, IIIT Allahabad, UP, India
E-mail: anupam69@gmail.com

Abstract— In the recent year gesture recognition has become the most intuitive and effective communication technique for human interaction with machines. In this paper we are going to work on hand gesture recognition and interpret the meaning of it from video sequences. Our work takes place in following three phases: 1. Hand Detection & Tracking 2. Feature extraction 3. Gesture recognition. We have started proposed work with first step as applying hand tracking and hand detection algorithm to track hand motion and to extract position of the hand. Trajectory based features are being drawn out from hand and used for recognition process and hidden markov model is being design for each gesture for gesture recognition. Hidden Markov Model is basically a powerful statistical tool to model generative sequences. Our method is being tested on our own data set of 16 gestures and the average recognition rate we have got is 91%. With proposed methodology gives the better recognition results compare with the traditional approaches such as PCA, ANN, SVM, DTW and many more.

Index Terms— Haar Cascade, Adaboost Algorithm, Hidden Markov Model, Douglas Peucker Algorithm.

I. INTRODUCTION

In recent years human and machine communication plays an important role to develop various fields like security, industries, hospitals, automation. In our daily routine work gestures are helpful for the handicapped peoples and old peoples to do their work without taking help of others and dumb person can easily express their feelings and emotion with others. Gesture recognition has played an important role in the field of research and become the important part of HCI (human computer interaction) [1]. In past years people are using touchable devices for giving input to the system like keyboard, mouse, and remote control so on [12] but now days they got change to untouchable devices we called it gesture. In market other than gesture recognition based devices are available but due to uncomfortness such devices are not popular.

There are many approaches for it some used wearable devices like gloves, helmets and for depth images they are using sensing rings. But human don't want to wear all these devices as it creates irritation to the people many times to wear them. So vision based gesture recognition is being used to overcome from this difficulty. We can use hand gesture for robot control, augmented reality, in mobile phones, in home tasks etc. Gestures can be takes place in two classes' first one static and second is dynamic. On the basis of this recognition takes place in two ways like static hand posture recognition, dynamic hand gesture recognition. In spite of static hand gestures [18, 19] dynamic hand gestures gives good communication results with the help of motion information and some gestures are based on RGB-D sensors [17]. Approaches for hand gesture recognition are classified as vision based and data gloves based [1, 12, 13]. Glove based approach is being employ for the purpose of gesture recognition with the support of data gloves. Proposed work is based on hand movements because hand is suitable for movement in direction easily comparatively to other body parts. Many real life applications takes place with the help of dynamic hand gestures like sign language recognition and for controlling the computer and many more.

In this paper our work is totally concentrated on dynamic hand gestures recognition which is based on the hand movements. Tracking takes place to handle the dynamic hand gesture. Dynamic hand gesture is based on four features velocity, shape, orientation and location. Hand motion can be known as points sequence with centroid of person's hand performing the gestures. We trained gesture using HMM and creates a model for each gesture. Hand gesture recognition takes place using those trained gestures. Speech processing, Signal processing also takes place using HMM. Here markov model is used for getting information of the documents.

II. STATE OF ART

In previous years detection of hand gesture takes place with the help of mechanical devices to get information about the hand position and orientation for example data gloves. Vision- based gesture recognition replaces the use of wearable devices. It is more users friendly. It is on high demands for the dynamic hand gesture recognition. Basically there are three modules of gesture recognition pre-processing, tracking and recognition. For the tracking purpose there are many methods given by the researchers such as Viola Jones based classifier, it gives more robust

results for pattern detection [8]. Haar cascade classifier is more robust to detect in noise and all. Many researchers give different approaches for the work like PCA, HMM, SVM etc. The approach given by Chen-Chiung Hsieh et. al is based on MHI (motion history image) [3]. In this method silhouettes can be used to get the image of a person in a single template form. Four features are used by them to recognize the gesture they are velocity, position, shape and orientation. Work takes place on real time platforms also.

The approach given by Yang Zhong et. al is based on HMM [4]. Skin color segmentation is used here for computing the hand image sequence. Features used by them are hand position, size, shape, velocity. After that data aligning algorithm is being applied to align feature vector sequences for the purpose for training. Then HMM model is being implemented for each gesture and results are accurate and effective.

S.M. Shitole et. al [5] implemented a technique based on three methods PCA, Pruning and ANN. Here the work is done in two ways first on the basis of color marker and second on the basis of non color marker. The system which is using color marker PCA algorithm is applied and system which doesn't use color marker Pruning and ANN are used. In past years people are using speakers operated with the help of multiple devices like keyboard, mouse and so on but Charade et. al [12] uses hand gestures for operating all this and all the presentations. By doing this literature survey we see that the work done by the researchers proves that hand gestures can be practically used in various applications like power point control, vlc player control, in virtual avatars and many more.

III. PROPOSED METHODOLOGY

There are three stages for the completion of our work as given in figure:

Fig. 1. Stages of method use

On the basis of literature survey we decide to work with the Viola Jones algorithm and then we used the Camshift algorithm [6, 20] for the purpose of tracking. For extracting the feature we used Douglas Peucker algorithm [7] so that we can able to find out the critical points by which we can be able to find out the angle by the movements of the hand. We start by setting the first frame taken by the camera as a background image and then after detecting the hand we apply the background elimination. At the last stage for finding the hand we use the Viola Jones detector and the Viola Jones region. By utilizing the detector and tracker we are able to avoid the

ingeminate search of the same frame. When we got any action image of the hand we locked it until the other scheme takes place.

We have train a cascade of features we use it for finding and classifying the object. There are some features which we had taken in cascade so the window will be chosen and classified; the object is being seeked only if it sustains all the features which is taken in the cascade. It represents that most of the windows will not stay for long time and they can dismissed in a short time. By this we can search out the whole image. For making the classification even more robust overlapping windows is then moved and resized after some pattern and then tested against the cascade again. Sometimes we face a question about the image with maximum probability is really an object or not because of in excess overlapping in some particular part of image. For filtering out matches with too much overlaps we apply another threshold value.

The threshold value is totally depends upon the quality of the cascade means the features we have taken. After getting the region of interest, we applied the background elimination algorithm to eliminate the background. Then we applied Camshift algorithm [6, 20] to keep track on the notion of hand.

After getting the track on hand then there is task to get trajectory of motion. We took centre of the ellipse that bound the hand and mark its trajectory. Now we get a sequence of very large points but there will be error due to webcam used and due to the Camshift algorithm. To reduce that error and to reduce memory we used Douglas Peucker algorithm it will not take that points whose Euclidean distance are less than some epsilon and we only take changes into account

Our work is going to pursue in three modules first one is hand detection and tracking, second is features extraction and third one is gestures recognition. We have taken 16 dynamic hand gestures in our work which includes numbers 0 to 9, circle, x, y, l, rectangle, waving hands.

In Viola Jones we basically use only 5 types of feature. Every feature is being computed as the sum of addition and subtraction of rectangular pixel that summation is being compared by the threshold value is being computed when the training takes place. It is seen that the small number of features can be used to create good classifier [8].
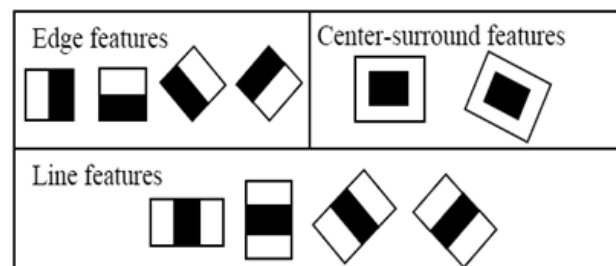

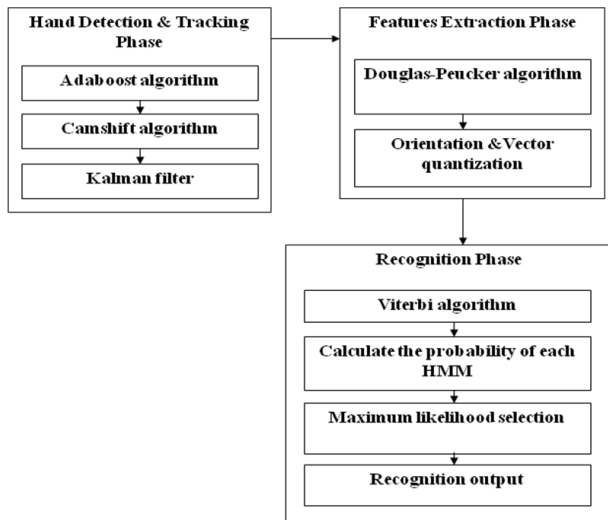Fig. 2. Haar like features by Viola Jones [8].

Fig. 3. three modules of proposed methodology.

Hand detection and tracking: In this first module we have to find out the Haar like features [14] consist of black and white jointed rectangles. Mathematically formula for Haar like feature is given by equation (1):

$$h(x) = Sum_{\,black\,rec\tan gle}\,(pixel\ grey\ level) -$$
$$Sum_{\,white\,rec\tan gle}\,(pixel\ grey\ level) \quad (1)$$

We make the use of integral images for finding the rectangle of Haar like feature with the help of integral images. The images holding the Haar like features are being examined with the support of sub-window for the detection of hand. Weak classifier can be computed by using Haar like features. Mathematically it is written as equation (2). Where fj(x) represents the weak classifier.

$$f_j(x) = \begin{cases} 1\ if\ p_j h_j(x) < p_j \theta_j \\ 0\ otherwise \end{cases} \quad (2)$$

Where p_j exhibit inequality sign's direction, threshold is equal to θ and sub-window is given by x. Now for improving the accuracy we use Adaboost algorithm. Adaboost algorithm is being completed in following steps:

Step 1: It starts with a weights uniform distribution.

Step 2: It takes weak classifier from learning algorithm.

Step 3: Increase the weights on the training examples those were misclassified

Step 4: repeat step 1 to 3.

Step 5: At each epochs there is a weak classifier which marks linear combination of it.

With the support of Haar like features we got the accurate results when the position of hand is in vertical direction, and missed in rest of the direction the problem can sort out by using Camshift algorithm [6, 20]. And noise is reduced with the help of kalman filter [9, 20].

Steps for Camshift algorithm:

Step 1. Select the search windows initial location.

Step 2. Calculate mean value with in the window.

Step 3. At calculated mean value in step 2 we centre the search window.

Step 4. Repeat step 2 & 3 for some epochs.

Step 5. Now we have to compute zeroth moment have to set the size of search window equal to zeroth moment function.

Step 6. Repeat step 4 & 5 till the pre set threshold value is more than mean value moves.
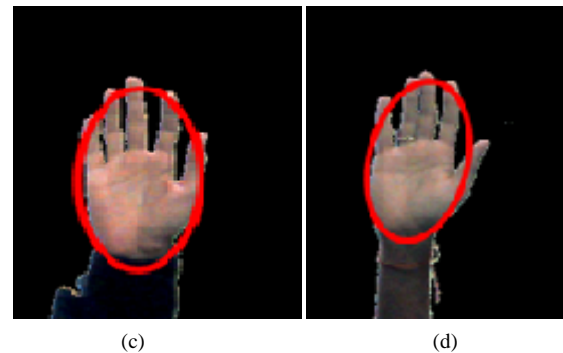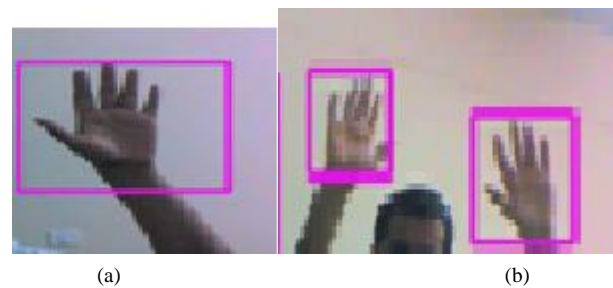

Fig. 4. skin color segmentation in HSV color space.

(a)   (b)

(c)   (d)

Fig. 5. outcomes of hand detection and tracking

Feature Extraction:

Feature extraction process is based on two modules first one is trajectory parameter. In this we take the center point of the hand and in the sequence we join these points to every frame. Center point of hand can be calculated with the help of moments [15] of pixels in hand's regions. Which is given by the equation (4):

$$M_{ij} = \sum_{X,Y} X^i Y^j I(X,Y) \quad (3)$$

Where X and Y gives us the range of the hand region and I(X, Y) represents the pixel value of image. So the center point of hand is given by the equation (5):

$$X_c = M_{10} / M_{00}, \; Y_c = M_{01} / M_{00} \tag{4}$$

We have lots of frames but all of them don't match the sequence which we got for the gesture. Second is the trajectory approximation when we are moving our hand for gesture recognition we see that the movement of hand is not so fast so it does not change frame to frame too much. So we can take only few frames for gesture recognition as there is no need to take all the frames. Douglas Peucker algorithm is being used for observing points of key trajectory. So here we got two points starting point and the ending point and having a threshold value so in between the value which comes close to the threshold value we discard it and which is under threshold value we consider it in the sequence and make a smooth curve with it. The codeword input is being calculated with the help of angle and angle can be calculated by using the equation (6)

$$Angle = \tan^{-1}(Y2 - Y1 / X2 - X1) \tag{5}$$
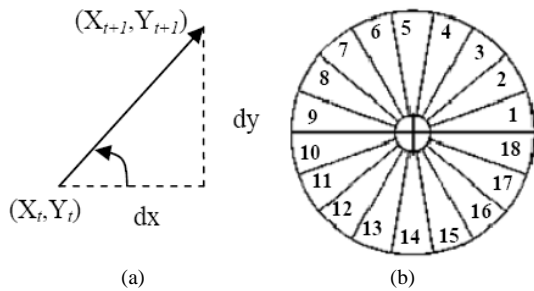

(a)   (b)

Fig. 6. slope point can be figure out in (4a), (4b).the set of codeword's [4].

Here we divide the angle to quantize them by 20 degree in 18 different codeword's. We have given 25 code wards as our input data for recognition process of HMM.

Recognition

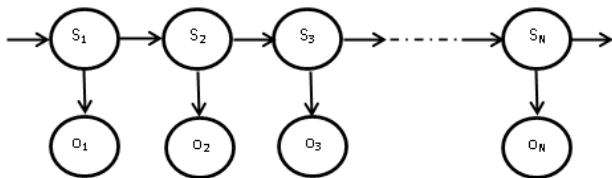Recognition process for dynamic hand gesture is done by using Hidden Markov Model.

Fig. 7. HMM

To understand the method used in the thesis first we have to understand the basic HMM model. In the above Fig. (7) S_t for t=1..N are representing the hidden states of HMM and the o_t for t=1..N are the observations.

The observations are given to the HMM on the basis of the features we have retrieved from each frame of the video. So it will be right to say that t = 1..N are just the N continued frames of a video

So the model will be:

$S_1....S_N \; \varepsilon \; \{1....M\}$ where M represents the number of states.

$o_1....o_N \; \varepsilon$ (discrete, real) showing that the observation can be discrete or real

$$O = (o_1....o_N)$$

So the joint distribution on all the variables will be

$$P(o_1 .... o_N S_1 .... S_N) = p(S_1)p(o_1|S_1)\prod_{t=2}^{N} p(S_t|S_{t-1})p(o_t|S_t)$$
$$\text{eq. 6}$$

The above stated eq. 6 represents the factorization which the graph model shown in Fig. 7 corresponds.

The HMM parameters in the above equation are -

Transition probability: $A(i,j) = P(S_{t+1} = j|S_t = i)$ for i, j ε {1....M}.   eq. 7

Emission probability: $B_j(O_t) = P(O_t|S_t = i)$ for i ε {1....M}.   eq. 8

Initial probability: $\pi_i = P(S_1 = i)$ for i ε {1....M}.
eq. 9

The HMM model is represented as (A, B, π) where A, B and π are described in eq7, eq8 and eq9.
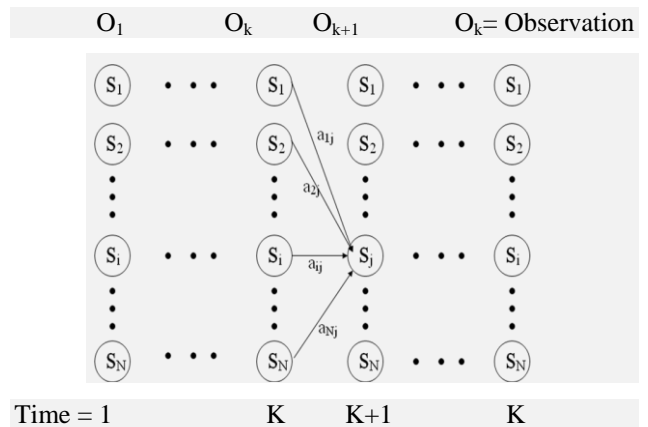

Fig. 8. Trellis representation of an HMM

Fig. (8) Showing the trellis diagram where the individual features detected in the video are used to find the interactive features and these interactive features are used as the observations O for the model (A, B, π).

Interactive features are observations for the HMM i.e. we came to know about the interaction between the persons with the help of features which are being calculated by taking the individual features. These individual features are separately calculated for different gestures and then using these individual features interactive features are generated which are then used as the observation for the HMM. The HMM works on two phases; first is the training phase and second is the recognition. The Baum-Welch algorithm is used for the training phase and the maximum likelihood is found for the recognition of the hand gesture.

When we are having some observations and from them we have to find out a model (A, B, π) we use Baum-Welch algorithm for its solution. By this we came to

know that we have given the observations of the gestures and by this we create a model (A, B, π) which maximizes the probability of given observations. By this we can train different gesture with the help of Baum-Welch training algorithm. The Baum-Welch algorithm makes use of forward - backward algorithm.

The **forward–backward algorithm** is known as dynamic programming implementation algorithm. It is implemented as a part of hidden markov model. The values of hidden variables of HMM required estimation. Posterior probabilities of these variables are being calculated by using forward-backward algorithm. Let the hidden variables be $S_k \varepsilon \{S_1, \ldots, S_t\}$, and transition/mass function with values $o_{1:t} := o_1, \ldots, o_t$, then this algorithm evaluates the probability distribution $P(S_k|o_{1:t})$. Two steps of recursion or initialization is being taken by the algorithm. In first step we take the forward values and in next step we take backward values. Due to this reason we called it forward-backward algorithm. This algorithm is used in speech processing here it creates HMM modeling takes place for large amount of data. The Hand gesture recognition can be known as finding the maximum likelihood of given model (A, B, π) and the observations $o_{1:t} := o_1, \ldots, o_t$. The maximum likelihood value for the model (A, B, π) and observation $o_{1,t}$ is given by $P(O|\lambda)$. Hand recognition takes place by calculating the maximum likelihood $P(O|\lambda)$ for all the model $\lambda(A,B,\pi)$ given the observations $o_{1:t} := o_1, \ldots, o_t$. The model having highest likelihood is taken as model to which input gesture belongs. In this model we can see the output of the state but cant able to see the state directly. By the word hidden we came to cognize that here we don't cognize the parameters of the model we just have sequence through which model has passed. When we got the trajectory after completing tracking algorithm, with the help of HMM we calculate the probability for every gesture by taking features which are extracted. Hidden markov model [10, 15, 16] is being implemented for solving the three problems which takes place for completing the markov process

a. Decoding: More likely parse sequence is found here. We employ Viterbi algorithm to overcome from this problem [10, 11].

b. Evaluating: compute the likelihood that the sequence is generated by a model. We employ Forward and Backward algorithm [10] for overcome this problem.

c. Learning: Maximization of the output probability of bring forth the symbol sequence. Baum-Welch algorithm [10] is being implemented for solving this problem.

After the completion of the training process for each gesture we are able to recognize the gesture which corresponds to maximum likelihood of 16 HMM model by using viterbi algorithm. Figure (5) shows the time sequence of image modeling in hidden markov model. Where f represents the frame and HMM1 to HMM 9 and so on are models for each gesture.
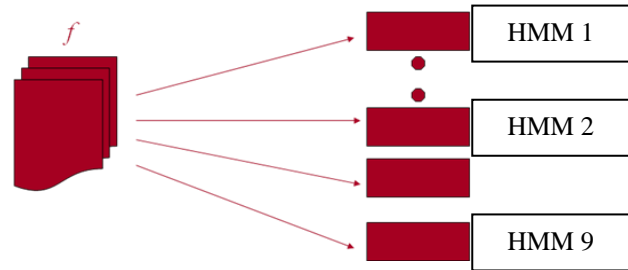

Fig. 9. HMM design model for each gesture.

IV. EXPERIMENTAL RESULTS

2 GB. Input is taken with the help of webcam. We have taken 16 gestures for the work. Accuracy rate is totally different for each gesture. In our work we have taken 160 video sequences for making cascade, we use 40 for training and 20 for testing. Video sequences are having positive and negative samples. Now we train each gesture using HMM and design HMM model for each gesture and test on real time. The whole result of the recognition is given in a table 1 with the help of percentage. The system will able to track and recognize the following gestures:

- User drawing number 0 with the help of hand movement in the air. Same thing happen with all other numbers from 1-9.
- User made waving hand as hand waving.
- User made x, y, z, circle, rectangle in the air with the help of hand movements.

There are some gestures here in which the motion of the hands is same so it becomes difficult to separate those gestures. But our algorithm works on it quietly and resolves this problem. Image representation of few of the gesture from all the gestures we have taken.
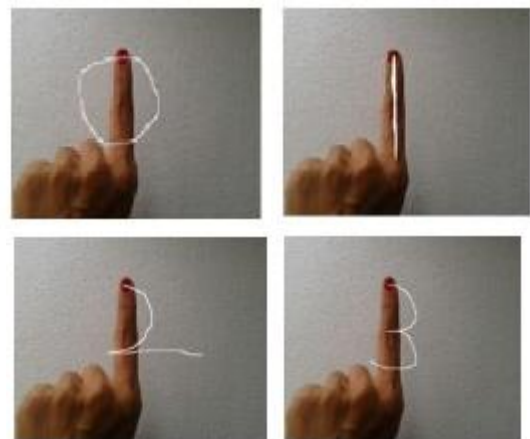

Fig. 10. images of dynamic hand gestures for number.

In table 1 we have compared our methodology with other method for few gestures. We see there that our method gives us more accuracy rate as compare to both two methods. So the recognition rate for Pruning is 84% [5], for ANN its 86% [5] and for our methodology average recognition rate is 94% for those five gestures and 91% for all gestures.

Table 1. Result of classifier

Gestures	Correct rate % for Pruning method	Correct rate % for ANN method	Correct rate % for Our method
0	90	90	94
1	80	80	92
2	85	90	96
3	90	90	90
4	ND	ND	86
5	ND	ND	94
6	ND	ND	94
7	ND	ND	98
8	ND	ND	86
9	ND	ND	90
Y	ND	ND	84
Circle	ND	ND	92
Rectangle	ND	ND	78
Waving hands	ND	ND	90
X	ND	ND	94
L	75	80	98

With the help of this table we can able to make comparison between the accuracy rates of our methodology with other methodologies. In this table ND represents the work is not done on these gestures by given methods.
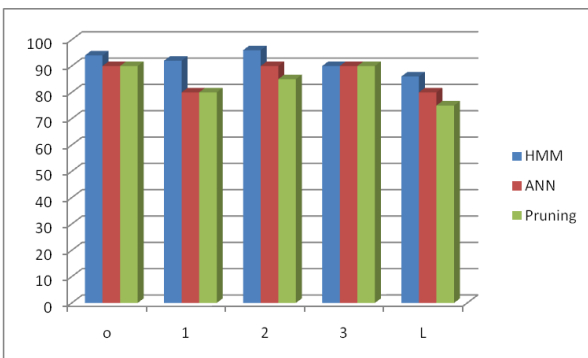

Fig. 11. Column chart of gesture recognition for 5 gestures with different method.

In figure 11 we see the column chart showing the recognition rate of 5 gestures using different method and make us help to make a comparison between the results of HMM with other two methods. In bottom we have 5 gestures and in left site there is the% recognition rate.

In figure 12 we see the % recognition rate for 16 gestures which I have taken for our work. Many of them are giving good results and one of them jus giving the accuracy of 75 %. In the figure below are gestures and in left % recognition rate.

Our results can be best described in the terms of confusion matrix. Confusion matrix is formed by detecting true and false matches of gesture. A confusion matrix is a table that reports the number of *false positives*, *false negatives*, *true positives*, and *true negatives*.

Confusion matrix represented by column chart in Fig. 4.4 shows a good result depicting 0.7 and above true positives for gesture recognition. Higher the true positives better is the performance of the algorithm.
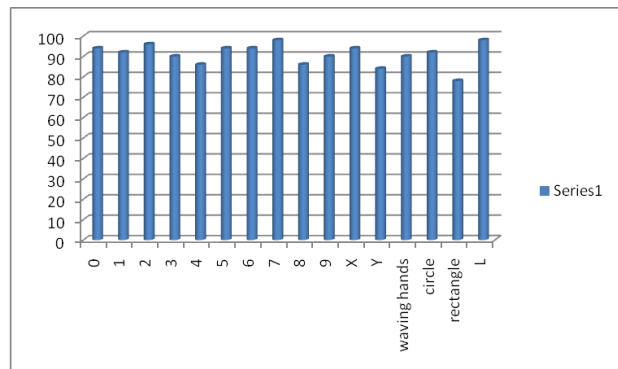

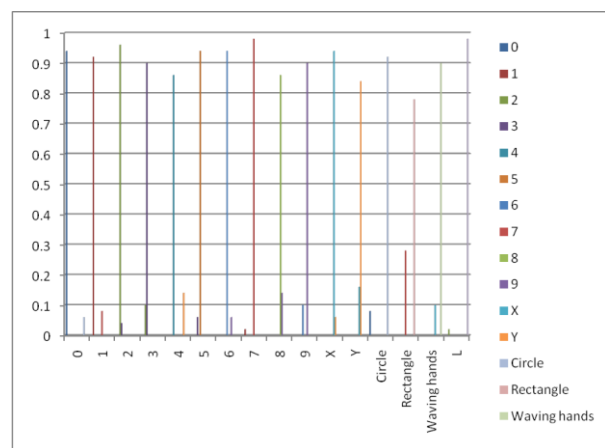Fig. 12. Recognition rate % for 16 gestures

Fig. 13. Column chart of the confusion matrix

V. APPLICATIONS

We configure our work with real time world activities. Here we control our system with the help of gesture each gesture used here for different – different task we map each gesture with some application of the system some of them are given below:

Gesture	Work in real time
0	Open chrome window
9	Open the vlc player
4	Increase volume of vlc player
6	Open the office

As the given above are very few examples of gestures mapped with real life application there are many more gesture mapped with some application of the system.

VI. CONCLUSIONS & FUTURE SCOPE

In our work we have define a method for recognizing ten types of different hand gestures. The whole implementation is done by taking the most interactive

features of the hand. With the help of these features we are able to give the best result in gesture recognition. The overall features we have taken for each gesture help us to recognize the gesture easily and accurately. The accuracy rate of recognition of the gesture is high as compared to previous results. We are able to reduce the confusion problem which takes place between some few gestures so we are able to short out this problem by our work.

In order to further enhance the system performance there are certain research issues that need to be addressed. First we will extend this thesis work for video having cluttered background and illumination on variant background with different detection and tracking algorithm. Second the number of gestures can be increased. Third we will do this work for 3d image.

REFERENCES

[1] W. Du and H. Li, "Vision based gesture recognition system with single camera," IEEE 5th International Conference on ICSP, Beijing, vol. 2, pp. 1351-1357, 2000.

[2] Gray Bradski and Adrian Kaehler, Learning OpenCV, O'Reilly Media, Inc.

[3] Chen-Chiung Hsieh, Dung-Hua Liou and David Lee, "A Real Time Hand Gesture Recognition System using Motion History Image," IEEE 2nd International Conference on Singal Processing Systems, Dalian, vol. 2, pp. 394-398, 2010.

[4] Yang Zhong, Li Yi, Chen Weidong and Zheng Yang, "Dynamic Hand Gesture Recognition using hidden Markov models," IEEE 7th International Conference on ICCSE, Melbourne, pp. 360-365, 2012.

[5] S.M. Shitole, S.B. Patil and S.P. Narote, "Dynamic Hand Gesture Recognition using PCA, Pruning and ANN," International Journal of Computer Applications, New York, vol.74, 2013.

[6] S.M. Nadgeri, S.D Sawarkar and A.D Gawande, "Hand Gesture Recognition using Camshift Algorithm," IEEE 3rd International Conference on ICETET, Goa, pp. 37-41, 2010.

[7] Kittasil Silanon and Nikom Suvonvorn, "Thai Alphabet Recognition from Hand Motion Trajectory using HMM," International Journal of Computer and Electrical Engineering, vol.4, 2012.

[8] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 511-518, 2001.

[9] M.S.M Asaari and S.A Suandi, "Hand Gesture Tracking System using Adaptive Kalman Filter," IEEE 10th International Conference on ISDA, Cairo, pp 166-171, 2010.

[10] Lawrence R "A tutorial on hidden markov models and selected applications in speech recognition," Proceedings of IEEE, vol. 77, pp. 257-286, 1989.

[11] G.D. Forney, The viterbi algorithm, proceeding of the IEEE pp. 268-278.

[12] T. Baudel, M. Baudouin-Lafon and Charade, "Remote Control of Objects using Free-Hand Gestures," Comm. ACM, New York vol. 36, pp. 28–35. 1993

[13] D.J. Sturman and D. Zeltzer, "A survey of glove-based input," IEEE conference on Computer Graphics and Applications," vol. 14, pp. 30-39, 1994.

[14] Qing Chen, Nicolas D. Georganas and Emil M. Petriu, "Vision Based Hand Gesture Recognition using Haar-like Features," IEEE conference on Instrumentation and Measurement Technology, Warsaw, pp. 1-6, 2007.

[15] L.W. Campbell, D.A. Becker, A. Azarbayejani, A.F. Bobick and A. Plentland, "Invariant Features for 3-D Gesture Recognition," IEEE Second International Conference on Automatic Face and Gesture Recognition, Killington, pp. 157-162, 1996.

[16] J. Schlenzig, E. Hunter and R. Jain, "Recursive Identification of Gesture Inputers using Hidden Markov Models", IEEE Second Workshop on Applications of Computer Vision, Sarasota, pp. 187–194, 1994.

[17] Jose Manuel Palacious, Carlos Sagiies, Eduardo Montijano and Sergio Liorente, "Human-Computer Interaction Based on Hand Gesture Using RGB-D Sensors," Journal on Sensors, vol. 13, pp. 11842-11860, 2013.

[18] S. S. Rautaray and A. Agrawal, "Real Time Hand Gesture Recognition System for Dynamic Applications," International Journal of UbiComp, vol. 3(1), pp. 21-31, Jan 2012.

[19] S. S. Rautaray and A. Agrawal, "Vision Based Hand Gesture Recognition for Human Computer Interaction: A Survey," Artificial Intelligence Review, Published Online: 06 November, 2012.

[20] J. C. Peng, L. Z. Gu and J. B. Su, "The Hand Tracking for Humanoid Robot using Camshift Algorithm and Kalman Filter", Journal of Shanghai Jiaotong University, 2006

Authors' Profiles

Varsha Dixit has completed M.Tech in the Department of Information Technology at Indian Institute of Information Technology, Allahabad, India. Her current research interests include Human-Computer Interactions, User Interface Design and Computer Vision Image Processing. She has published her paper in Springer Proceedings. Currently she is working as Assistant Professor in ITM University Gwalior.

Anupam Agrawal is presently working as Professor & course co-coordinator of HCI (Information Technology)at Indian Institute of Information Technology Allahabad (IIIT-A). Before joining IIIT-A in the year 2000, he was working as scientist 'D' at DEAL, DRDO, Govt. of India, Dehradun. He received his M.Tech. degree in Computer Science and Engineering from Indian Institute of Technology Madras, Chennai and Ph.D. degree in Information Technology from Indian Institute of Information Technology Allahabad (in association with Indian Institute of Technology, Roorkee). He was a postdoctoral researcher at the Department of Computer Science & Technology, University of Bedfordshire (UK) during which he had contributed significantly in two major European projects. His research interests include Computer Vision, Image Processing, Visual Computing, Soft-Computing and Human-Computer Interaction. He has more than 75 publications related to these areas in various international journals and conference proceedings, and has co-authored one book. He is on the review board for various international journals including IEEE, Springer, MDPI, Taylor & Francis and Elsevier. He is currently serving as a Principal Investigator of an international (Indo-UK) Project. He is a member of ACM (USA), senior member of IEEE (USA) and a fellow of IETE (India). He is also serving as Chairman of the ACM Chapter at IIIT-A.