Static Hand Gesture Recognition using an Android Device

Tejashri J. Joshi
M.Tech. Student
Vishwakarma Institute of Technology,
Pune

Shiva Kumar
Domain Lead
Persistent Systems Limited,
Pune.

N. Z. Tarapore
Assistant Professor
Vishwakarma Institute of Technology,
Pune

Vivek Mohile
Consultant
Persistent Systems Limited,
Pune

ABSTRACT
The need to enhance communication between humans and computers has been instrumental in determining new communication models, and accordingly new ways of interacting with machines. A vision based Hand Gesture Recognition system can be useful to recognize hand gesture in air, with devices like camera equipped smart phones and cameras connected to computers. The fast improvement of smartphones during the last decade has been predominantly determined by interaction and visualization innovations. Despite the fact that touchscreens have significantly enhanced interaction technology, future smartphone clients will request more natural inputs, for example, free-hand association in 3D space. To extract the features of air gesture we used statistical technique which is Principal Component Analysis (PCA). The recognition approach used in this paper is based on Support Vector Machine (SVM). Proposed Hand Gesture System is location and orientation invariant. All the processes to recognize the hand gesture are done on the device. This approach can be easily adapted to a real time system.

Keywords
Hand Gesture Recognition, Android, Principal Component Analysis, Support Vector Machine, Pattern Recognition, Mobile Computer Vision

1. INTRODUCTION
Mobile computing has entered into a golden age, with a significant increase in the use of smartphones. Smartphones are typically powered with multi-core CPUs which make them powerful computing devices. Most of these devices are also equipped with multi-core GPUs like Adreno420, NVidia Tegra, etc. The high configuration on a portable mobile phone devices helps in faster execution of heavy computational tasks, such as rendering realistic 3D animations in high definition resolution, and augmented reality. This improvement in hardware performance focuses on mobile computing, enabling people to perform operations with their mobile phones faster instead using a computer. Additionally, every smartphone today is equipped with two cameras' (rear and back) which is helpful for capturing picture's, videos, etc.

In the past, Computer Vision software was limited to desktop computers. Advancement in the mobile hardware makes it possible to utilize computer vision software's on mobile platform.

1.1 Gesture recognition
In computer Science and language technology, gesture recognition aims at interpreting human gestures using mathematical algorithms. Gesture implies movement of the body parts which has specific meaning/message used to communicate between sender and receiver. Gesture Recognition is a way for machines to understand human body language, thus bridging a gap between machines and humans. Gesture may involve any one of the body parts or combination of one or many of them. Hand gesture has great impact of human communication. It has been the most common way of communication. Hand gesture involves static gesture and dynamic hand movements. Static gesture is nothing but hand posture, for example jamming the thumb and forefinger to show 'ok' symbol. In dynamic hand gesture movement of the hand is involved like waving of hand. Types of gestures recognition, then subtypes of the vision based gesture recognition system which are appearance based technique and 3D model based techniques are explained in the survey [1]

2. RELATED WORK
For the interaction with mobile currently we rely on the touch input but this limits the expressiveness of the input. Recently some models are designed based on image processing and machine learning to recognize the hand gesture on the resource constrained devices like smartphones, smartwatches. Here we review the literature that expands the input space from touchscreen to the areas around mobile and to help deaf and dumb persons to communicate with each other.

AndroSpell [2] uses android camera to grab the image and further preprocessing feature extraction and classification is done on device. For Classification machine learning algorithm such as decision tree and neural network are used. It recognizes the hand poses with accuracy of 97%. This system is beneficial only for finger spelling recognition. Researched designed Hand Gesture Recognition system [3] based on client server architecture. The image is acquired using the IPWebCam app and all the captured frames are sent to the server. Preprocessing stages like edge detection thinning the object to reduce the noise is done and tokens of the hand images are created. These tokens are used as a feature and passed to the neural network for classification. It recognizes hand gesture with 77% accuracy. The main drawback of this system is client server model.

System based on Finger-Earth Mover's Distance [4] recognizes fingers using device of 3 GB RAM with 93% accuracy. This system is invariant to location, scale and orientation. Some system requires sufficient distance between the fingers for recognition but this system it is able to recognize the fingers which are connected to each other. Because of the short length of the thumb, system gets confused to recognize the gestures which involves thumb. This system works well using Kinect but using only RGB camera it is required wear a black band for segmentation. Recently machine learning based in-air Hand Gesture Recognition system [5] is proposed which uses multistage classification. Segmentation is done using simple thresholding technique. Depth classification forest estimates the depth of the image and frames in 15-50 cm range are forwarded. Shape Classification tree is used to classify 6 gesture classes as well as noise. Part classification forest classifies fingertips for pointing-like gestures. The overall accuracy of this system is 83%. As with all the vision based system using RGB camera has issue of ambient light this system also has same issue. Segmentation technique in used in the system is sensitive. It will not work well for dynamic gesture because the nature of classifier is based on per pixel value.

3. PROPOSED SYSTEM

The algorithm proposed in this paper has three main stages- Preprocessing, Feature Extraction and Classification stage. The preprocessing stage prepares the input image, obtained from camera for further processing. Feature extraction stage uses the preprocessed image and extracts feature using Principal Component Analysis (PCA). PCA finds principal components that retain 90% of the image data. This essentially finds the direction in which maximum data lays, in-turn reducing the dimension of the image by more than 9%. This is very useful for classification in real-time. The training is performed using examples of different gestures shown in images. The system runs the images through preprocessing stages and extracts relevant features from images. Like PCA Support Vector Machine also needed setup procedure. These features are used to train multiclass Support Vector Machine to properly classify the input image.

Indirect method of multiclass SVM constructs the several binary SVM and combines the output for the final class. As mention in [6], there are three methods in this category- one-against-all, one-against-one and Directed Acyclic Graph (DAG SVM). These methods lead to existence of the unclassified regions. To resolve this problem Decision Tree based multiclass SVM (DTSVM) is used. For N-class problem, it calculates N-1 hyper-planes which are less than conventional method. As training proceeds, the amount of data required for processing becomes smaller and hence results in shorter training time.

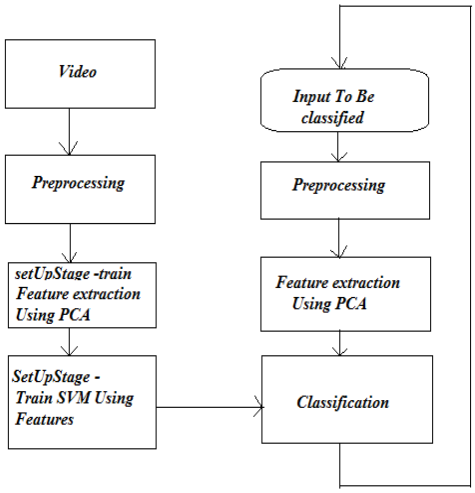

Figure 1:SystemArchitechture

Implemented Hand Gesture Recognition System can be train not only for the hand gesture but also for the various patterns. There are few constraints of the system. Black background is required for recognition .User need to perform gestures by showing hand from the wrist.

3.1 Preprocessing

Preprocessing plays a vital role in Gesture Recognition System. It is used to segment the region of interest from background, remove image noise, and normalize the area of interest and other operations that will help to reduce the representation of the image. By using a reduced feature set that still contains all the required data which is used to distinguish the gestures, it's simple to represent pattern and classify gestures. Outcome of this is faster classification using less memory.

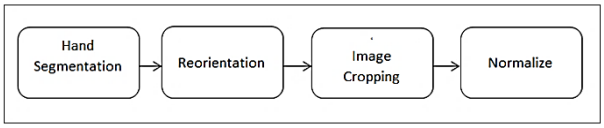

Figure 2: Preprocessing Stages

3.1.1 Hand segmentation

The aim of image segmentation is to make simpler image and/or change the representation of the image into something which is more meaningful and easier to analyze. In current system to segment hand from background, we used thresholding which is simplest method of segmentation. In this method each pixel in the image is replaced by black pixel if image intensity$I_{i,j}$is less than threshold or a white pixel if image intensity is greater than that threshold.

$$dst(x,y) = \begin{cases} \max val & if\ src(x,y) > threshold \\ 0 & otherwise \end{cases} \quad (1)$$

Here $dst(x,y)$ is value of destination matrix and $src(x,y)$ is the value of source image at pixel location (x, y) and max value is 255.

3.1.2 Reorientation

Size and orientation have too much impact on pattern matching. All the training images are straight and of different sizes. For user it's difficult to give straight images, and this leads to orientation problem. To solve orientation problem, the orientation angle needs to be found. As mentioned in [7],

before we can find the orientation angle we need to find out the center of the image. So the center of the image is going to be found and then orientation method will be applied to the center. Angle of rotation is the angle between axis of least second moment and the center of the image, which is found using equation (2).

$$2\theta = \arctan\left(\frac{2M_{11}}{M_{20}-M_{02}}\right) \quad (2)$$

Here M_{11} is central moment indicate the center of the mass and $M_{20,}M_{02}$ gives the axis of least second moment.

When we use center of the mass of data as the center of rotation in the source image, small amount of data is lost and it leads to reduction in accuracy. The image is rotated along the center of the image.

Figure 3 represents the angle of orientation between center of mass and axis of least second moment and the center of image at which image can be rotate.

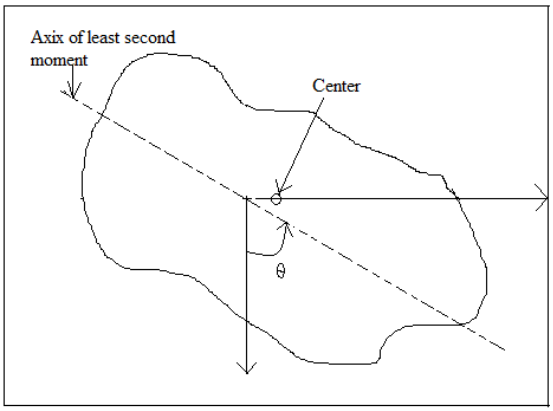

Figure 3: Angle of orientation

3.1.3 Image cropping

Working on area of interest instead of working on the complete image makes computation easier. It results in location invariant system. So it is necessary to crop the region of interest from image. To crop the image we used horizontal and vertical profiling to reduce the matrix into vector. Row and column of the matrix is treated as asset of 1 d vectors and performing the operation of sum of all the vectors until single row\column is obtained. As the background is black, we found first and last nonzero value from both vectors reduced from the matrix. From the start and end values we calculated subset of original image and get a cropped image.

3.1.4 Normalize

Generally PCA is used for dimensionality reduction, now-a-days it's also used for feature extraction. In some algorithms it is must to perform related preprocessing step called whitening. If we are training on images, the raw input is not so useful, since adjacent pixel values are highly correlated. Zero mean and unit variance helps in maximizing the covariance between the two poses and minimizing the covariance between similar images. It also helps to reduce light effect.

3.2 Feature extraction

In this stage we extract the features from the hand image. We extract all the features of particular gesture that represent how it differs from the other gestures. These features are then used in the recognition stage to classify gestures. In the proposed system Principal Component Analysis (PCA) is used to extract the features of the hand. PCA is commonly used for data reduction here it is used to extract the features. PCA minimizes the number of dimensions by compressing the image without losing too much information. In case of and gesture recognition, it is sufficient to analyze and store the difference between the individual hand images rather to process on whole hand image.

As mention in [9, 10] PCA is used for face recognition, sign language recognition respectively. Here we refer it to extract the hand features using PCA. PCA steps are

Principal Component Analysis

1: Create a set S of m training images; transfer each image into vector of size 1*5400

$$S = \{\tau_1, \tau_2, \tau_3, \ldots., \tau_m\}$$

2: Compute the mean image

$$\varphi = \frac{1}{m}\sum_{i=1}^{m}\tau_i$$

3: Find the difference between each training image and the mean image

$$\Phi_i = \tau_i - \Psi$$

4: Obtain the covariance matrix C

$$C = \frac{1}{m}\sum_{i=1}^{m}\phi_i\phi_i^T = \frac{1}{m}AA^T \; where \; A = [\phi_1, \phi_2, \phi_3, \ldots., \phi_m]$$

5: Compute Eigen Vectors of covariance matrix.

$$u_i = Av_i$$

where v_i are the eigen vectors of AA^T and u_i is the Hand Space

6: Project the image into space using

$$p_k = u^T(\Gamma_k - \varphi) \; where \; k = 1,2,3,4 \ldots., m$$

The feature of each image in the training dataset is p_k. Feature vector of all the training images are computed. These feature vectors will be further utilized to train the classifier.

Figure 4: Original image, output after Segmentation, reorientation, image cropping

3.3 Classification

Initially, extracted features are tried to classify using minimum Euclidean Distance between the feature vectors of training image and test image. Euclidean distance measures the correlation between quantitative, continuous variables. It is not suitable for ordinal data, where preferences are listed according to rank instead of according to actual value. It leads to reduce the accuracy.

To overcome the drawback of the Euclidean Distance, we tried simple classifier K-means Clustering. All the features of the same class are combined to for a cluster. Minimum distance between the cluster centers and the feature vector of the test image is calculated and predicted the class of the test image. It is not robust to outliers so it reduces the accuracy.

To avoid all the drawbacks of above classifiers we decided to use for SVM. Feature extracted using PCA are classified using a multiclass Support Vector Machine (SVM). The SVM is supervised learning models with associated learning algorithms which will analyze data and recognize patterns, if is used for classification and regression analysis. Basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary classifier. SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a margin that is as wide as possible

As mentioned in [6] there are indirect methods of the multiclass SVM which are one-against-one, one-against-all, Directed Acyclic Graph (DAG) SVM. All method uses binary SVM to solve the multi-class problem. These methods generates unclassified region. It degrades the generalization ability. To overcome this problem, we used decision tree based multiclass SVM. High generalization ability can be maintained by separating most separable classes at the top node of decision tree. For N class problems, N-1 hyper-planes are to be calculated which are less than that of indirect methods. There are four types of methods involved in the decision tree based SVM which are explained in [8]. Further comparison between different multiclass SVM methods can be found at [8].

As mentioned in the [8] steps to generate tree

Decision Tree based SVM

1: Calculate the class centers c_i (1… N) by

$$c_i = \frac{1}{|X_i|}\sum_{x \in X_i} x$$

And distance between the class i and j,$d_{i,j}(i, j = 1,2,..,N)$, by

$$d_{ij}(d_{ji}) = \| c_i - c_j \|$$

Here $X_i$is a set of training data included in class I and $|X_i|$ is the number of elements included in X_i

2: For the classes find the smallest value of $d_{i,j}$ for each class. Namely, the smallest value of class I,

$$l_i = min_{j=1,\ldots,N, j \neq i} d_{i,j},$$

Here let the associated two classes belong to the same clusters.

3: Repeat step 2 N-2 times to merge into two clusters.

4: Calculate a hyper-plane which separates the clusters generated in step 3.

5: If the separated cluster in step 4 has N' (>2) classes regard the classes as belonging to different clusters and repeat step 3 N'-2 times bad go to Step 4.If N' = 2, calculate a hyper-plane which separates the two classes.

Using this algorithm the tree is generated. At each node some classes are separated from remaining classes. Test image belongs to the class indicated by the leaf node.

4. EXPERIMENTAL RESULTS

In the previous section, system structure was described, preprocessing stages, feature extraction stages were elaborated. The System is trained for five gestures indicate one to five fingers. The Training set consists of gestures of the 7 people. The training part of SVM and PCA is done on the desktop machine. Average Image, Feature Space generated using PCA is utilized on the mobile device to extract the features of the test image.

The process of training multiclass SVM includes the training of several binary SVM's. All the SVM models at each node of the tree are stored in the XML files. Trained models of PCA and SVM are kept with the application. So after installation of the application all trained models are copied into the memory card of the device for further processing.

The cost of android smartphone is less than Apple, Microsoft Windows and BlackBerry. So we used android device. The testing device has 5.0MP of back camera with the Dual-core 1.2 GHz processor and 1 GB RAM. Our system is implemented in Android using OpenCV library.

The gestures were tested by performing them directly in front of the mobile camera. By performing all the image processing on device we get 10 frames per seconds.

We conducted our experiments using three different techniques SVM, Euclidean distance and the K-means clustering. Following figures 5, 6, 7 show experimental results in the form of confusion matrix. From these confusion matrices we have calculated accuracy for SVM, Euclidean distance and the K-means clustering techniques. The mean accuracy using clustering is 45%, using Euclidean Distance is 72.4% and using SVM is 97.6%. Out of the three techniques combination of PCA and SVM gives the best accuracy. Hence we have implemented our system using PCA and SVM.

	Predicted Class					
	Gesture	1	2	3	4	5
Actual Class	1	**100**	0	0	0	0
	2	0	**100**	0	0	0
	3	0	0	**96**	0	4
	4	0	0	0	**100**	0
	5	0	0	8	0	**92**

Figure 5: Confusion matrix using combination of PCA and SVM

	Predicted Classes					
	Gesture	1	2	3	4	5
Actual Class	1	**100**	0	0	0	0
	2	18	**76**	6	0	0
	3	6	18	**67**	9	3
	4	4	16	13	**67**	0
	5	0	11	22	15	**52**

Figure 6: Confusion matrix using combination of PCA and Euclidian Distance

	Predicted Classes					
	Gesture	1	2	3	4	5
Actual Class	1	**28**	24	22	26	0
	2	11	**42**	26	16	5
	3	30	13	**28**	23	3
	4	0	2	5	**85**	8
	5	0	16	7	25	**42**

Figure 7: Confusion matrix using combination of PCA and K-means Clustering

Figure 8 summarizes the results of the static hand gesture recognition System.

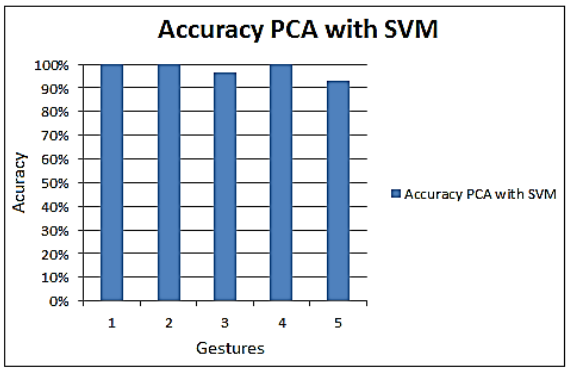

Figure 8: Positive recognition rates with 5 gestures

Comparison of experimental results between Euclidean Distance, SVM and clustering is shown in figure 9.

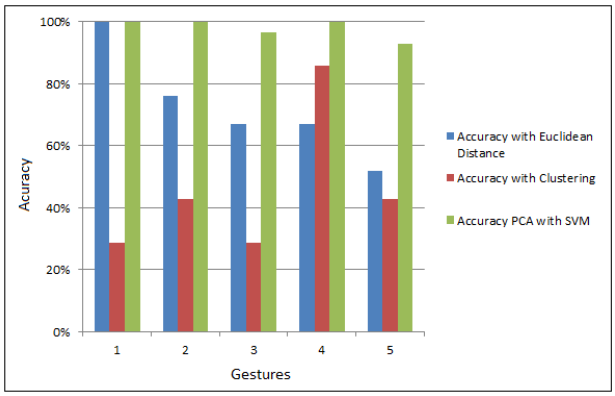

Figure 9: Comparison of experimental results using different classification techniques

5. CONCLUSION AND FUTURE WORK

In this paper, a new approach was proposed to recognize static gesture on the resource constrained devices like smartphones. The focus of this research is applying PCA for feature extraction, with low computation. For the classification Decision tree based SVM can be used to avoid the unclassified region generated in the conventional methods of multiclass SVM. From study of SVM it can be conclude that SVM achieves excellent result with very few samples. We achieved 97.6% accuracy for recognition of static hand poses.

From the study of skin color segmentation methods its conclude that there is still problem of segmentation of hand using normal RGB camera in skin color background as well as change in the light condition. Proposed Gesture recognition system recognizes the gesture in constrained environment like black background. So, further work should focus on hand segmentation method on the resource constrained devices with varying light condition and skin color background. As we increase the gesture set the tree based SVM classifier will increase the computation cost. So to minimize the computational cost of the tree based SVM may include specialized multiclass SVM classifier.

6. ACKNOWLEDGEMENT

I express my warm thanks to Prof. (Dr.) R. M. Jalnekar (Director VIT, Pune) and Prof. S. B. Karthick (HOD Department of Computer Engineering, VIT, Pune) for encouraging and granting me an opportunity to work with Persistent System Limited, Pune.

I would also like to thank my project manager Aarti Desai and my entire team including Rashmi Singh, Pranjal Shingane, and Sagar Gandhi at Persistent System Limited, Pune for providing me with the required facilities and an environment conducive for my M. Tech. project.

7. REFERENCES

[1] Rajeshri Rahul Itkarkar, Anil Kumar Nandy, "A Study of Vision Based Hand Gesture Recognition for Human Machine Interaction", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 12 (December 2014).

[2] Elhenawy, I., & Khamiss, A. (2014). The design and implementation of mobile Arabic fingerspelling recognition system. International Journal of Computer Science and Network Security, 14(2), 149-155.

[3] Saxena, A., Jain, D. K., & Singhal, A. (2014, April). Hand Gesture Recognition Using an Android Device. In Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on (pp. 819-822). IEEE.

[4] Ren, Z., Yuan, J., Meng, J., & Zhang, Z. (2013). Robust part-based hand gesture recognition using kinect sensor. Multimedia, IEEE Transactions on,15(5), 1110-1120.

[5] Jie Song, Gabor Soros, Fabrizio Pece, Sean Ryan Fanello, Shahram Izadi, Cem Keskin, Otmar Hilliges, "In air Gestures Around Unmodified Mobile Devices", UIST'14, October 5–8, 2014, Honolulu, HI, USA.

[6] Soman, K. P., Loganathan, R., & Ajay, V. (2009). machine learning with SVM and other kernel methods. PHI Learning Pvt. Ltd..

[7] Ghasemzadeh, A. (2012). Comparison of Feature Based Fingerspelling Recognition Algorithms (Doctoral dissertation, Eastern Mediterranean University).

[8] Takahashi, F., & Abe, S. (2002, November). Decision-tree-based multiclass support vector machines. In Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on (Vol. 3, pp. 1418-1422). IEEE.

[9] Smith, L. I. (2002). A tutorial on principal components analysis, February 2002.URL http://www. cs. otago. ac. nz/cosc453/student_tutorials/principal_components. pdf.(URL accessed on November 27, 2002).

[10] Ankita Saxena,Deepak Kumar Jain,Ananya Singhal, "Sign Language Recognition Using Principal Component Analysis", Fourth International Conference on Communication Systems and Network Technologies 2014.

[11] Ilan Steinberg, Tomer M. London, Dotan Di Castro, "Hand Gesture Recognition in Images and Video", center for communication and information technologies, March 2010.

[12] Dongseok Yang, Jong-Kuk Lim, Younggeun Choi, "Early Childhood Education by Hand Gesture Recognition using a Smartphone based Robot", The 23rd IEEE International Symposium on Robot and Human Interactive Communication August 25-29, 2014. Edinburgh, Scotland, UK,

[13] Joyeeta Singha, Karen Das, " Indian Sign Language Recognition Using Eigen Value Weighted Euclidean Distance Based Classification Technique", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 2, 2013.

[14] Luís Tarrataca, André Coelho, João M.P. Cardoso, "The Current Feasibility of Gesture Recognition for a Smartphone using J2ME", Conference: Proceedings of the 2009 ACM Symposium on Applied Computing (SAC), Honolulu, Hawaii, USA, March 9-12, 2009.

[15] Joyeeta Singha, Karen Das, " Indian Sign Language Recognition Using Eigen Value Weighted Euclidean Distance Based Classification Technique", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 2, 2013.

[16] Qing Chen, Nicolas D. Georganas, Emil M. Petriu, "Real-time Vision-based Hand Gesture Recognition Using Haar-like Features", Instrumentation and Measurement Technology Conference – IMTC 2007, Warsaw, Poland, May 1-3, 2007

[17] Joyeeta Singha, Karen Das, " Indian Sign Language Recognition Using Eigen Value Weighted Euclidean Distance Based Classification Technique", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 2, 2013.

[18] Rafiqul Zaman Khan, Noor Adnan Ibraheem, "Gesture Algorithms Based on Geometric Features Extraction and Recognition", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, ISSN: 2277 128X , November 2013.

[19] Kakumanu, P., Makrogiannis, S., & Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. Pattern recognition, 40(3), 1106-1122.