AN APPRAISAL OF AUTOMATED HAND GESTURE RECOGNITION TECHNIQUES

M. Jamil, M. K. Mahmood* and Y. D. Khan

Department of Computer Science, School of Science and Technology, University of Management and Technology, Lahore
*Department of Mathematics, University of the Punjab, Lahore
Corresponding Author's e-mail: khalid.math@pu.edu.pk

ABSTRACT: Hand gesture segmentation for identification and prediction analysis has been a prevalent topic in the community of the researchers working in the field of image processing and computer vision. Hand gestures were used by humans to signify various expressions, moods and actions. The advent of intelligent devices has made hand gesture recognition even more interesting. Many researchers have worked on intelligent interfaces that use hand gestures for its operations. Various state of art techniques used for hand gesture recognition has been critically analyzed. A method using Hahn moments and neural networks has been proposed. A subsequent comparative study has been drawn in terms of performance of the proposed and other competitive techniques. It was observed that the proposed technique was quite accurate and efficient.

Keywords: Hand gesture recognition, Hahn moments, Receiver operating characteristics, image processing.

(Received 15-09-2015 Accepted 23-11-2016)

INTRODUCTION

In the current era our lives are dependent on digital machines such as, computers. The influence of technology is so overwhelming that it is unimaginable to live without it. Humans feel compelled to use machines for the purpose of work, communication and even entertainment. To make it more convenient and comfortable, technology requires better and more comfortable means of interaction with machines. Some years ago machine-man interaction was not as interactive as to day because of some limitations. Human Computer Interaction (HCI) has evolved from wired devices to wireless, due to which, machines have become easier to operate. In some cases automatic mechanisms are used such as intelligent cameras which are able to capture and perform operation based on human gestures requiring very little or no human interaction. Researchers have gained interest in vision based interaction as this is a more natural and informal way to communicate with computers (Nasri *et al*., 2015).

Typically hands are used to point, move or modify objects. Hands move in any direction but some hand gestures have special meanings like, saying hello or waving hands left or right to point towards a direction. Hand movement with single move is called posture and continuous motion of hands is called gesture (combination of different static posture with a specific sequence). Thus hand movement is a mean of non-verbal communication (Tan and Guo, 2012).

A great amount of work has been done to reduce the limitations that are involved in Human Computer Interaction (HCI). Several researchers had presented various solutions to the problem using different mathematical or statistical models (Khan *et al*., 2014 and Khan *et al*., 2012). Subsequently, many researchers have worked with different devices and instruments to recognize hand movements, using gloves or markers to segment gestures. But in these techniques user have to wear the device which is usually not comfortable. Moreover, researchers in the field of computer vision and machine learning have worked to make it more realistic. By using these advancements in technology, people who cannot communicate orally can also get benefit. Many techniques have been used to recognize hand gestures by using mathematical approaches. Some of these techniques have been incorporated into dedicated devices like, contour-based similarity (Saeed and Behrad, 2014) or Convex Shape Decomposition Method (Qin *et al*., 2014). Hidden Markov Model (HMM) has been a very popular technique for modeling and classifying dynamic gestures (Tanand Guo, 2012). The work of (Liu *et al*., 2008) represents human actions as a combination of the movement of the body part. They propose a representation that is portrayed by a blend of human body-part developments compared with specific activities. The polar space is utilized to explain the example of every human body-part development.

Spin-Images and Spatial-Temporal (ST) features are used by (Liu *et al*., 2008) with Fiedler Embedding Euclidian Space to identify the relationship among different actions. Numbers of deficiencies are observed in the paradigm given by (Liu *et al*., 2008). The kinematic features do not work well if an image is viewed from different angles. The reason is the exhaustive nature of their approach. Kinematic features require 30-40 seconds for each video, while the kinematic modes can oblige only upto 10 seconds against each video. The introducing

step is the slowest due to irrational memory usage and the iterative nature of the task. The study of (Nasri *et al.*, 2015) identifies the hand gesture of same contour-based images (CBSI). Hand shape information is used to detect hand movements. They extract the hand contour from frame and then calculate the center of the mass of hand. For CBSI matching, Scale-invariant features (SIFT) are used. Support Vector Machine (SVM) is used by (Nasri *et al.*, 2015) for real time hand recognition using novel Haar features, this paradigm is designed to work on skin color detection. Real-time hand gesture recognition from depth images has been discussed by (Qin *et al.*, 2014) using convex shape decomposition method. Gesture feature extraction for static gesture recognition given by (Hasan and Kareem, 2013) is not invariant to translation cases. Utilizing immaculate hand shape is insufficient for fruitful results. Recognizing human actions by learning and matching shape-motion prototype trees has been discussed by (Shao *et al.*, 2012). Human action segmentation and recognition via motion and shape analysis is described by (Guha and Ward, 2012). They use color intensity methods to segment actions by manually selecting a region. Gesture Recognition with Applications as shown by (Bilal *et al.*, 2013) could have issues with other objects in the scene that have a similar hue and saturation.

MATERIALS AND METHODS

To understand hand motions, diverse methodologies were utilized like vision based, multiple features, etc. The general gesture recognition process is elaborated as follows:

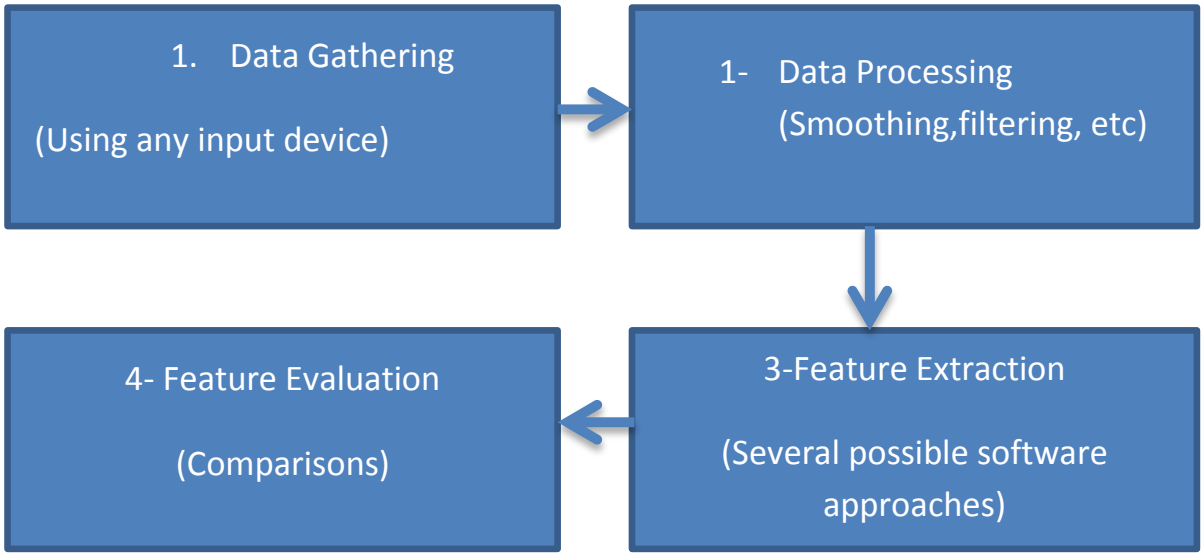

Figure-1Describes the general process of Hand Gesture Recognition

Spatial temporal and Spin-Image Features were used by (Liu *et al.*, 2008) for detecting actions. To ideally consolidate these peculiarities, they added to a structure that took into consideration learning of certain connections between distinctive classes of features in a principled way. The structure was focused around the idea of Fiedler Embedding. They connected K-means for bunching the list of features.

Contour-based images (CBSI), as shown by (Liu *et al.*, 2014), required images of size of $K * K$ pixels. The power of the $(i, j)^{th}$ pixel in CBSI is given as $CBSI(i, j) = Sim\ (C_i\ ,\ C_j)$ where C_i and C_j werethe clipping zone of the imagealong the i^{th} and j^{th} plots, and *Sim* was a measure of the intimacy of the limits. CBSI worked strongly even though the size, speed, and introductory area of the hand were varied. To represent and calculate the hand contouring, the following formula was used:

$$x_c = \frac{\sum_{i=1}^{N} x_i^I}{N} y_c = \frac{\sum_{i=1}^{N} y_i^I}{N}$$

Also,

$$x_i^c = x_i^I - x_c y_i^c = y_i^I - y_c$$

Where (x_c, y_c)were the centroid and (x_i, y_i)were arbitrary points.

TheCBSI method is based on the SIFT features of input image which were used for training and learning. The method proposed by (Hasan and Kareem, 2013) for hand detection scheme consisted of three steps:

1- Foreground segmentation
2- Palm localization
3- Hand segmentation.

Foreground F was extracted using the following equation:

$$F = \{(p,z_p) \mid z_p < z_0 + z_D\}$$

Where (p, z_p) denoted the pixel in the depth image at coordinate p with value z_p, z_0 was the minimal value of the depth image and z_D was the threshold.

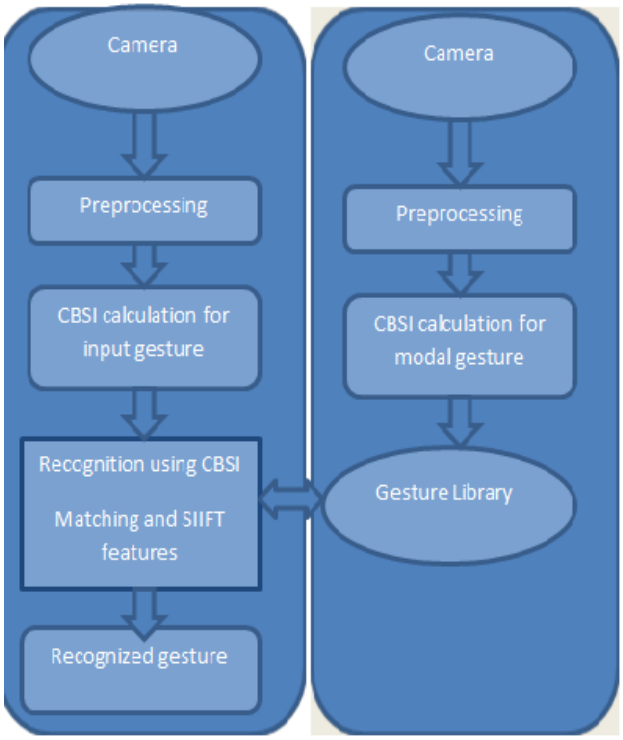

Figure-2: Block diagram of CBSI (Nasri *et al.*, 2015)

Following 7x7 surround mask was used for convolution filter:

-5	-6	-5.5	-5	-5.5	-6	-5
-6	-5	-2	0.5	-2	-5	-6
-5.5	-2	0.4	0.4	0.4	-2	-5.5
-5	-0.5	0.4	225	0.4	0.5	-5
-5.5	-2	0.4	0.4	0.4	-2	-5.5
-6	-5	-2	0.5	-2	-5	-6
-5	-6	-5.5	-5	-5.5	-6	-5

Figure-3A 7x7 surround mask used for convolution

Image filtered using the mask in Figure-3 was used to extract features of the image. The feature vector was collected by computing complex moments of the filtered image. Complex moment of order m was calculated by using following equation:

$$C_m = \int\int (x + i_y)^m \ \mu(x, y) dx\, dy$$

Where, $i = \sqrt{-1}$ and $\mu(x, y)$ was the intensity function of real image.

The hand picture of size $n \times m$ was figured by taking after mathematical equation at n^{th}-order complex minute (M_i). For the hand image of size $(n * m)$ the feature vector was computed using the mathematical equation

$$M_i = \sum_{x=0}^{n-1} \sum_{y=0}^{m-1} Z_n^i \, a(x,y)$$

Complex number $a(x, y)$ was expressed as $Z_n = X_n + \mathrm{i}\, Y_n$ with respect to a pixel position. Most of the time approximation of complex moments needed some extra computational time.

The formula was composed by combining the real and imaginary part as;

$$Z_n = R_n + \mathrm{i}\, I_n$$

By investigating the case Z_0 and Z_1, it was established that $R_0 = 1$.

$I_o = 0$, $R_1 = x$, $I_1 = y$ replacing the estimation of Z_n, $Z_n - 1$ and Z_1 yielded

$$R_1 = R_n - I_x - I_n - I_y$$
$$I_1 = I_n - I_x + R_n - I_y$$

Hand gesture recognition was applied to both posture and gesture by (Hasan and Haitham, 2013; Hasan and Kareem, 2013 and Hsieh and Liou, 2015). While (Liu and Jingen, 2008; Nasri *et al.*, 2015 and Hsieh and Liou, 2015) worked on dynamical hand movements.

Depth information of image was used by (Qin *et al.*, 2014) to detect the hand action. Shape decomposition and features were used to identify palm, skeleton and fingertips of hand. Depth images were working with different light intensities and camera-object distance problem had been reduced successfully.

Hand center of decomposing image was called as palm and other parts were known as fingertips.

Fingertips were defined by using following formula:

$$f_{tip} = \{f_j | \max dist(cut_i, f_j), f_j \in S_f\}$$
$$S_f = \text{Shape of fingers}$$
$$cut_i = \text{Cut between fingers}$$

Fingertips shape descriptors were used to find out the different features of hand structure from palm and these features were used in training.

Contour based information was used by (Hasan and Kareem, 2013) to extract features which were free from sizing and scaling issues. Sobel operator, edge detection and Laplacian filters were used in this technique. Edge was differentiating from color intensity in an image; same formula was used for contouring an image. Neural Network was used for classification of the images.

In a study carried out by (Rautaray and Agarwal, 2015) the authors gave a detailed summary for recognizing an action based on visual data and reported that 21% of the time, hand was utilized in actions as compare to other body parts. Contact gadgets, mechanical devices or ultra-sonic gloves help the users to interact with system. The work of (Chen and Dung, 2012) took a shot at novel features model by the following steps shown in Figure-5:

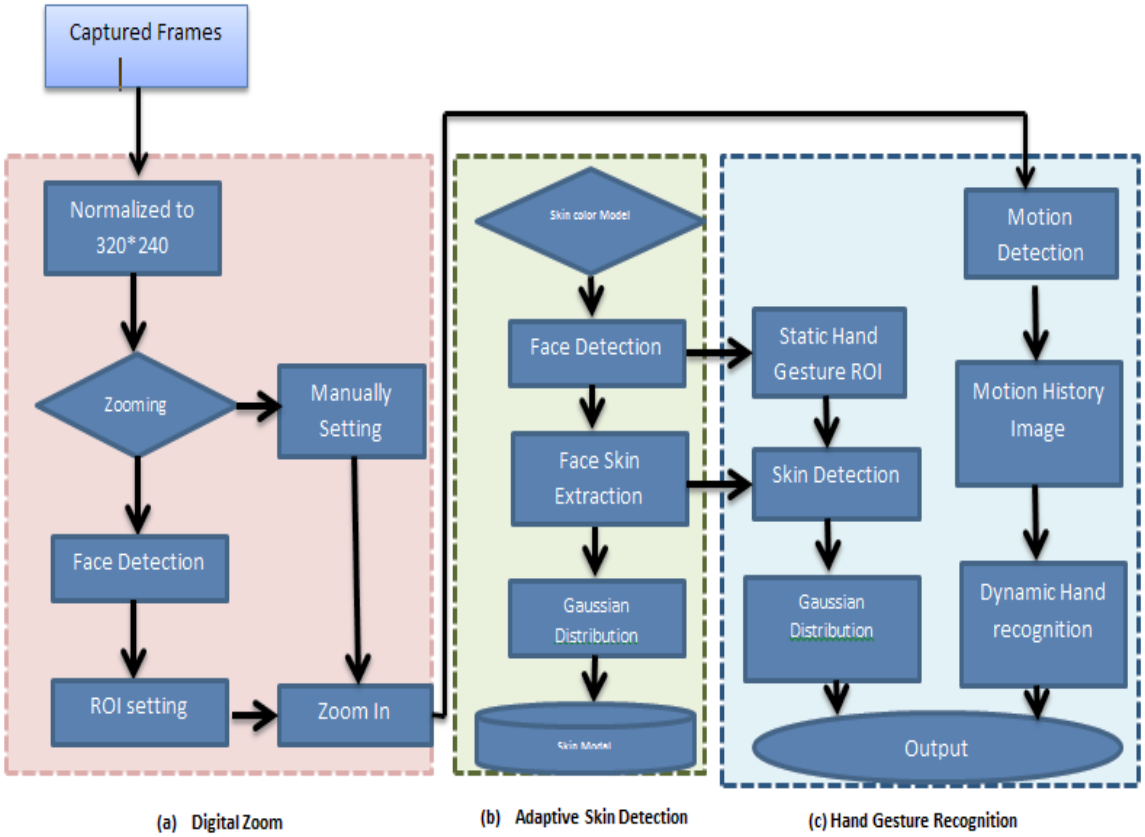

Figure-4 Flow Diagram that shows the three phases of the process

Firstly, dynamic zoom was performed using the following equation:

$ROI_1(x,y,w,h) = (face_{max}$ .x-5 X $face_{max}$.r, $face_{max}$ .y-5 X$face_{max}$ .r, 10 X $face_{max}$ .r, 10 X $face_{max}$ .r)

Where *x,y, w* and *h* were arbitrary points that defined the region of interest.

Secondly, Motion History Images (MHI) were utilized for progressively identifying movements. Constant movement data was utilized to upgrade the movement history picture using the following equation.

$$MHI(x,y)_t = MHI(x,y)_{t-1} + DF(x,y)_{t-1} - \propto$$

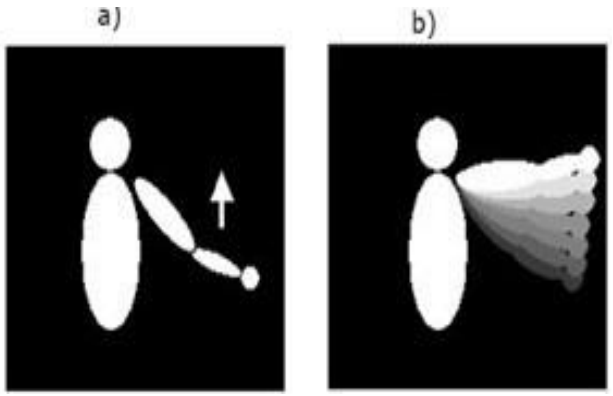

Figure 5. Motion History Image depicting the contours from previous frames

Curved shape decay system was employed for hand shape disintegration and representation as illustrated by (Qin *et al.*, 2014). Various preprocessing steps were applied as discussed by (Hasan and Kareem, 2013) which were based on image segmentation techniques such as edge identification and limit recognition.

Shape series of an object within certain set of frames described a specific action useful to classify it as shown by (Shao *et al.*, 2012). Action cycle could be detected from two methods; the first method took the history of changed color intensity values and other method calculated the difference between frames and generated a vector which was called motion gradient. Features were extracted from video by using its temporal behavior. Multi class support vector machine was used for classification.

Actions were represented by a well-defined structure of prototypes corresponding to the video frames (Jiang *et al.*, 2012). Region of interest was selected for obtaining feature descriptors. K-means clustering was used for action prototype learning and finally binary trees of prototypes were created. Every node of tree independently represented each prototype. In order to detect motion frame to frame difference was computed and maps were used to shape motion binary tree prototypes. Later, k-means clustering was used for

classification of input action. Results with moving camera using this technique are displayed in figure-6.

The proposed technique had the capability to extract vision based features as well as temporal features. The feature vector was derived using its geometrical attributes. Hahn moments were computed to extract the scale invariant features of each frame. Two dimensional Hahn moments were orthogonal moments that required a square matrix as a two dimensional input data. The Hahn polynomial of order *n* was given as

$$h_n^{\mu,\nu}(r,N) = (N+\upsilon-1)_n(N-1)_n \times \sum_{k=0}^{n}(-1)^k \frac{(-n)_k(-r)_k(2N+\mu+\upsilon-n-1)_k}{(N+\upsilon-1)_k(N-1)_k}\frac{1}{k!}$$

The above expression used the Pochhammer symbol generalized as;

$$(a)_k = a.(a+1)\dots(a+k-1)$$

And was simplified using the Gamma operator

$$(a)_k = \frac{\Gamma(a+k)}{\Gamma(a)}$$

The raw values of Hahn moments were usually scaled using a weighting function and square norm given as;

$$\widetilde{h_n^{\mu,\nu}}(r,N) = h_n^{\mu,\nu}(r,N)\sqrt{\frac{\rho(r)}{d_n^2}}, \qquad n = 0,1,\dots N-1$$

While ,

$$\rho(r) = \frac{\Gamma(\mu+r+\upsilon)\Gamma(\upsilon+r+1)(\mu+\upsilon+r+1)_N}{(\mu+\upsilon+2r+1)n!\,(N-r-1)!}$$

The orthogonal normalized Hahn for the two dimensional discrete data were computed using the following equation:

$$H_{ij} = \sum_{q=0}^{N-1}\sum_{p=0}^{N-1}\beta_{pq}\widetilde{h_i^{\mu,\nu}}(q,N)\widetilde{h_j^{\mu,\nu}}(p,N), \qquad m,n = 0,1,\dots N-1$$

Hahn moments up to order 3 were calculated. The obtained feature vector was used to train a neural network. After extraction of features from the training data, the feature vectors were fed to a back propagation neural network, as described in (Khan *et al.*, 2012). The neural network was sufficiently trained. The trained neural network was tested on test data to identify hand gestures. The Cambridge Hand Gesture Data set was used.

RESULTS AND DISCUSSION

The work of (Saeed and Behrad, 2014; Qin *et al.*, 2014; Tanand Guo, 2012 and Liu *et al*., 2008) used various stochastic and intelligent models to recognize hand gestures, based on visual features, whereas the contributions of (Nasri *et al.*, 2015 and Hasan and Kareem, 2015) pertained to congregation of spatial temporal features and characteristics. Some researchers also emphasized the need of a hybrid approach.Table-1 show a comparison of the best accuracy in results produced using various motion and/or shape related characteristics.

Table 1.Showing results of action recognition using shape-motion prototype.

Method	Motion only	Shape only	Joint shape and motion
Recognition Rate (%)	87.5	53.57	92.86

A lot of work was done on classification for images but it was challenging in case of videos when space and time was involved as shown in the work of (Guha and Ward, 2012). They introduced a method for sparse representation by using a dictionary that can retrain whenever new data is incorporated. Three possibilities were used for constructing dictionary i.e. shared, class specific and concatenated. In shared framework, a common dictionary was saved once, and used for all *k* classes having some common features. There were two matrices *X*, *Y*. *X* contained the corresponding sparse representation and *Y* contained the training samples descriptors.

In class-specific, *k* dictionaries there were k coefficients i.e.$\varphi_1, \varphi_2, \varphi_3, \dots, \varphi_k$which were trained per class. Benefit of this technique was that each class was independent but extra effort was required when new class is introduced into the system.

In all K-dictionaries were concatenated to form concatenated dictionaryφ_c. They used public data set to evaluate the proposed techniques on both action and facial expressions. Cuboids and LMP descriptors were used for grouping. Comparative performance for these descriptors was illustrated below:

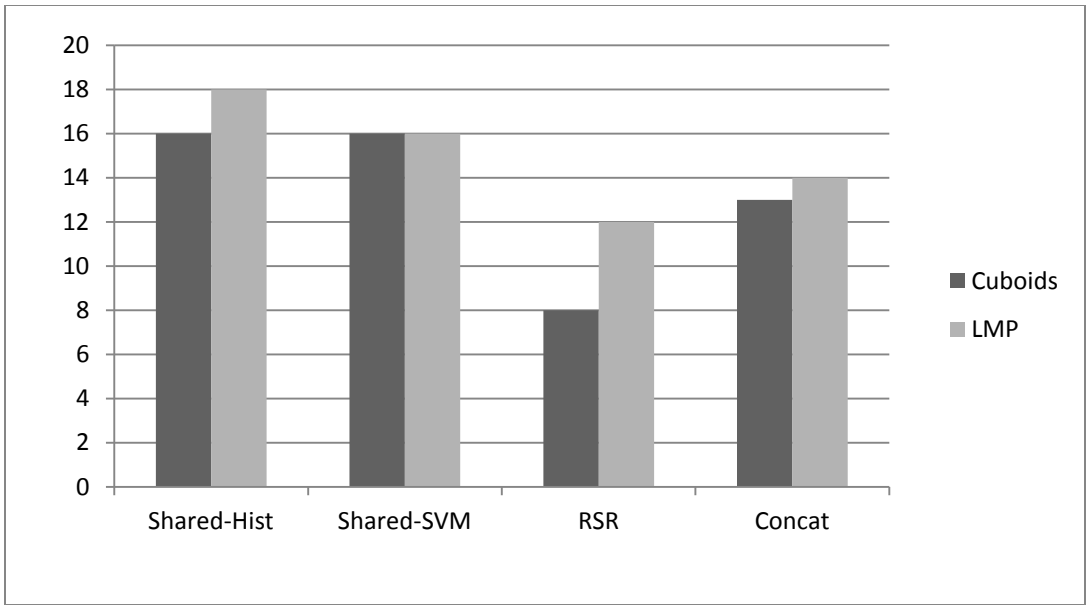

Figure-6Comparative analysis for action classification

The performance of the work of (Guha and Ward, 2012) was illustrated below for sparsity (k_i , $i = 1,2,3$) against shared, RSR and concatenated techniques:

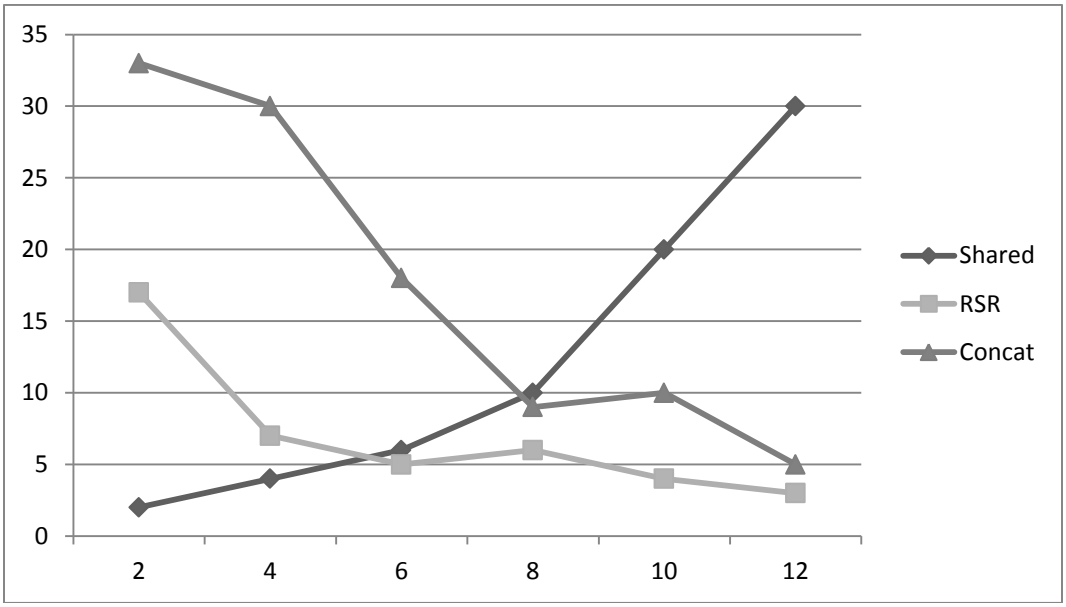

Figure-7Performance of shared, RSR and concatenated techniques.

(Bilal *et al.*, 2013) compared different techniques of action recognition using Hidden Markov Models (HMM). HMM is very useful from speech learning to word recognition. Different sign language recognition systems used HMM with accuracy from 82% to 94.8%.

The work of (Tan and Guo, 2011) reported on six section frames: the administrator, the picture securing module, hand situating module, 3D enrollment module, motion distinction module and the virtual working module. The work of (Tan and Guo, 2011) utilized current standards and Cam Shift to discover the target rapidly and precisely. They utilized position of focus and zone of shape to discover and recognize hand form.

Comparisons of some essential research analysis related to hand movement detection system were described in table2. Generally some specific actions were used for classification. An overall behavior of different techniques on these actions was illustrated.

Table 2 Showing evolution of gesture recognition technologies.

Year	Key Method	Application
2008	Fiedler Embedding	General in nature
2015	CBSI,SIFT	American Sign Language
2014	Neural Network	Action for HCI
2013	convex shape decomposition method	virtual interaction
2013	Sober Operator, Complex Moments	Speech Recognition
2015	SVM, novel Haar-like features	embedded systems interaction
2015	Vision based, Ada Boostalgorithm Hidden Markov Models	interaction medium for mouse and keyboard
2012	Motion history image	detection of human actions in an indoor environment
2012	hierarchical K-means clustering	Support dataset
2012	spatio-temporal descriptors, scale invariant	model human actions
2013	Hidden Markov model	Sign Language
2013	Cam Shift,	Virtual chemistry experiment platform,
2016	Hahn Moments and Neural Networks	Hand gestures

The Cambridge Hand Gesture Data set used by (Kim and Cipolla, 2009) is the most vastly used database for hand gestures techniques. This database was used to generate feature vectors for each image using the methodology described by (Nasri *et al.*, 2015; Bilal *et al.*, 2013; Feng and Yuan, 2013; Hasanand Kareem, 2015). The major classification models used were, Neural Networks (NN), Support Vector Machine (SVM), and Hidden Markov Model (HMM). Using these models for the same database, the images were classified. True Acceptance Rate (TAR) depicted the rate with which the system classified images of a certain class as belonging to its respective class. While, False Acceptance Rate (FAR) was the rate by which the system accepts an image as belonging to some other class. Receiver Operating Characteristics (ROC) distribution was plotted to represent the TAR and FAR ratios accordingly for the comparison of proposed and previous competitive techniques. Figure-8 shows the ROC graph.

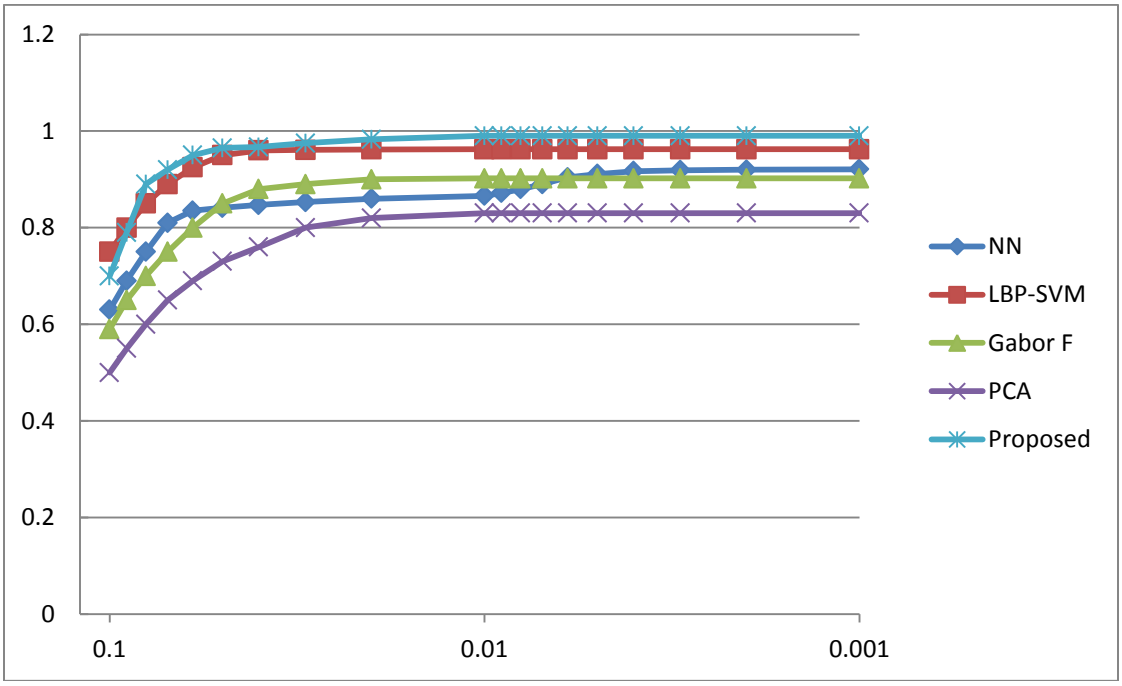

Figure-8. Receiver Operating Characteristics graph for competitive and proposed hand gesture recognition techniques.

It was observed from the ROC distributions that the area under the curve of SVM based technique was maximum, which means that the proposed technique was sufficiently accurate.

Similarly, the techniques illustrated by (Nasri *et al.*, 2015; Hasan and Kareem, 2015; Qin *et al.*, 2014; Hasan and Kareem, 2013 and Hsieh and Liou, 2012) and the proposed technique were evaluated using prescribed databases. For each generalized hand gesture the accuracy of the system was computed using a confusion matrix. The accuracy in percentage along with the number of samples used for each gesture is illustrated in table-3. This shows that the proposed technique's performance is efficient as well as accurate.

Table-3.Showing the accuracy of hand gesture recognition for each major class.

Reference #	Close (%)/total gestures	Left(%)/total gestures	Right(%)/total gestures	Up(%)/total gestures	Open (%)/total gestures
(Qin *et al.*, 2014)	96/127	94/215	93/242	91/233	96/146
(Hasan and Kareem, 2013)	77/9	80/10	78/10	60/10	50/8
(Hasan and Kareem, 2014)	80/50	75/45	60/50	90/50	65/50
(Nasri*et al.*, 2015)	80/60	75/60	86.7/60	78.3/60	71.6/60
(Hseih and Liou, 2015)	89.2/625	96/720	95.16/716	89/600	80/450
Proposed	91.2/625	90.6/500	92/512	91.8/600	90.6/450

Over the recent years, the utilization of common human hand signals for communication with supportive devices was a flourishing area of research. Different techniques were used for both static (posture) and dynamic (gesture) recognition. There were some limitations in essential systems that needed to be tackled and gave the degree of interest for future exploration.

During the recent decade various strategies for hand classifications and representations were proposed. Analytical study discussed in this paper expressed that the appearance based hand signal representations were more favored than the 3D based motion representations in the hand motion detecting and recognizing framework. Despite the fact that there were variety of features and examination productions available for both methods but the 3D model based representations were less favored because of complex nature of usage.

REFERENCES

Bilal, S., R. Akmeliawati, A.A Shafie and M.J.E. Salami (2013). Hidden Markov model for human to computer interaction: a study on human hand gesture recognition. *Artificial Intelligence Review*. *40*(4): 495-516.

Feng, K. P., and F. Yuan (2013). Static hand gesture recognition based on HOG characters and support vector machines. *Instrumentation and Measurement, Sensor Network and Automation. International Symposium:* 936-938.

Guha, T., and R.K. Ward (2012). Learning sparse representations for human action recognition. *Pattern Analysis and Machine Intelligence. IEEE Transactions.34*(8): 1576-1588.

Hasan, H., and S.A. Kareem (2014). Static hand gesture recognition using neural networks. *Artificial Intelligence Review*.*41*(2): 147-181.

Hasan, H. S. and S.B.A. Kareem (2013). Gesture feature extraction for static gesture recognition. Arabian Journal for Science and Engineering.38(12): 3349-3366.

Hsieh, C. C., and D.H. Liou (2015). Novel Haar features for real-time hand gesture recognition using SVM. *Journal of Real-Time Image Processing*. *10*(2): 357-370.

Jiang, Z., Z. Lin and L.S. Davis (2012). Recognizing human actions by learning and matching shape motion prototype trees. *Pattern Analysis and Machine Intelligence. IEEE Transactions*. *34*(3): 533-547.

Khan, Y. D., F. Ahmad and M.W. Anwar (2012). A neuro-cognitive approach for iris recognition using back propagation. *World Applied Sciences Journal*. *16*(5): 678-685.

Khan, Y. D., F. Ahmad and S.A. Khan (2014). Content-based image retrieval using extroverted semantics: a probabilistic approach. *Neural Computing and Applications*. *24*(7-8): 1735-1748.

Khan, Y. D., S.A Khan, F. Ahmad and S. Islam (2014). Iris recognition using image moments and *k*-means algorithm. *The Scientific World Journal*.

Kim, T. K., and R. Cipolla (2009). Canonical correlation analysis of video volume tensors for action categorization and detection. *Pattern Analysis and Machine Intelligence. IEEE Transactions. 31*(8): 1415-1428.

Liu, J., A. Syed and M. Shah (2008). Recognizing human actions using multiple features. *Computer Vision and Pattern Recognition. IEEE Conference:* 1-8.

Nasri, S., A. Behrad, and F. Razzazi (2015). A novel approach for dynamic hand gesture recognition using contour-based similarity images. *International Journal of Computer Mathematics*. *92*(4): 662-685.

Qin, S.,X. Zhu, Y. Yang, and Y. Jiang (2014). Real-time hand gesture recognition from depth images using convex shape decomposition method. *Journal of Signal Processing Systems*.*74*(1): 47-58.

Rautaray, S. S., and A. Agrawal (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*.*43*(1): 1-54.

Shao, L., L. Ji, Y. Liu, and J.Zhang (2012). Human action segmentation and recognition via motion and shape analysis.*Pattern Recognition Letters*. *33*(4): 438-445.

Tan, T. D., and Z. M. Guo (2011). Research of hand positioning and gesture recognition based on binocular vision. *IEEE International Symposium:* 311-315.