Dynamic Hand Gesture Recognition using Hidden Markov Model by Microsoft Kinect Sensor

Archana Ghotkar
Pune Institute of Computer Technology, Pune, India

Pujashree Vidap
Pune Institute of Computer Technology, Pune, India

Kshitish Deo
Pune Institute of Computer Technology, Pune,India

ABSTRACT
Hand gesture recognition is one of the leading applications of human computer interaction. With diversity of applications of hand gesture recognition, sign language interpretation is the most demanding application. In this paper, dynamic hand gesture recognition for few subset of Indian sign language recognition was considered. The use of depth camera such as Kinect sensor gave skeleton information of signer body. After detailed study of dynamic ISL vocabulary with reference to skeleton joint information, angle has identified as a feature with reference to two moving hand. Here, real time video has been captured and gesture was recognized using Hidden Markov Model (HMM). Ten state HMM model was designed and normalized angle feature of dynamic sign was being observed. Maximum likelihood probability symbol was considered as a recognized gesture. Algorithm has been tested on ISL 20 dynamic signs of total 800 training set of four persons and achieved 89.25% average accuracy.

General Terms
Human Computer Interaction, Pattern Recognition, Machine Learning

Keywords
Indian Sign Language, Dynamic hand gesture recognition, Hidden Markov Model

1. INTRODUCTION
In India, the main mode of communication between the deaf and the outside world is Indian Sign Language. Even after standardization and usage of ISL by the deaf, each time a deaf tries to communicate, he needs a human interpreter to make other people understand his language, who might not be available all the time.

This paper presents use of computer interface to communicate with deaf people. Even though, while performing a gesture, there are mainly hand gestures, other expressions like lip movements and facial expressions must also be taken into account. All ISL alphabet and number signs are composed of hand postures. Approximately 80% of ISL sign words are composed of hand gestures [1]. Considering the major role of hand gestures in the ISL interpretations, the paper focuses on signs made with the hands.

There are broadly two approaches to deal with gesture recognition, i) hardware based ie. use of gloves/sensors and ii) vision based. Vision based approach is found more suitable and practical as compared to hardware based approach [1]. Vision based processing can be divided into broadly two types: 2D web camera and 3D stereo camera. Problems faced by the 2D cameras are: complex algorithms to handle segmentation of colour, lack of depth information and brightness factor. With the advancement of sensor technology in computer vision, Microsoft Kinect sensor is widely used to deal with dynamic hand gesture recognition problem. Various researchers are using Microsoft Kinect sensor for SL recognition system [4]. Its advantages are i) provide depth and skeleton information, ii) ease of use iii) no constraints of background and iv) quite inexpensive compared to other stereo cameras.

Few of the models used for recognition purpose such as Neural Networks, Fuzzy logic and HMMs. This paper presents a system to recognize hand gestures with Kinect camera. Pattern of hand movement is analyzed using Hidden Markov Model. This gesture is divided into states. Output symbols are extracted from the gestures. These symbols become parameters for the Hidden Markov Model (HMM). In this paper, dynamic hand gesture recognition for ISL word using Kinect camera and HMM is presented.

2. LITERATURE SURVEY
Research is going on across the world in this area. Research in computerization of American Sign Language [2, 18], Japanese SL [3], and Swedish SL [4] have reached substantial level. In India, research on ISL interpretation started late due to lack of standardization and very less work is going on at present to convert ISL into sound language. In recognition phase, different methods were used like SVM [5, 6], KNN [7], HMM [9, 10] and DTW [8]. Most of the researchers were using data glove for continuous SLR. Typical phases include: segmentation of the hands based on the colour of gloves or the skin, then extraction of the features mainly the location, orientation and velocity of motion. In India, Kishore and Kumar [16] worked on ISL word recognition using fuzzy logic and achieved 96% accuracy. They used the extraction of features based on the colour.

There exist many reported research projects related to learning and recognizing visual behavior. However due to its recent introduction to the vision community, only a small number have been reported which use Hidden Markov Models. HMM has been traditionally used as tool for speech recognition tool. Recent researches have begun relating the speech variations to visual gestures. We summarize a few of the interesting works related to the paper. Kalin and Jonas trained the system with HMM model. 51 signs were analyzed which achieved 89.7% average recognition accuracy [4].Moni M. A. and Ali have analyzed various techniques and approaches in gesture recognition for sign language recognition using HMM [13]. They have provided an overview of HMM and its use in vision based applications, working in two stages that of image capturing and processing using cameras and the second stage for identifying and learning models has eliminated the need of previously used sensor embedded equipment such as gloves for tracking of a gesture. T. E. Starner has employed HMM in 1995 in identifying the American Sign Language. On similar grounds authors Gaus Y. et al. have successfully recognized the Malaysian Sign Language [12]. It consists of skin segmentation procedure throughout frames and feature extraction by centroids, hand distances and orientation has

been used, gesture paths define the hand trajectory. Kalaman filters have been used by researchers to identify overlapping hand-head and hand-hand regions. Elmezain et al. have quantized features form spatio-temporal trajectories into code words [9]. They have used a novel method of tracking the gesture by using 3D depth map along with colour information, this helps at separating the same colour at different surfaces in a complex background. In order to separate continuous gestures a special zero codeword is defined, using the start and end points of meaningful gestures the viterbi algorithm is employed. In [10] the authors have used the LRB topology along with forward algorithm to achieve the best performance. With recognition rate of 95.87% Arabic numbers have been identified. Shrivastav R [11] has used OpenCV image processing library to perform the isolation of gesture frames, the entire process form per-processing to testing. In coordination with this processing, Baum-Welch algorithm and LRB topology with forward algorithm is applied for recognition.

With the advent of sensor based camera Kinect, due toits benefits over the traditionally used vision based cameras, research orientation was shifted towards use of Kinect [17]. Kalin and Jonas [4] have developed educational signing game based on isolated sign recognition of Swedish sign language using Microsoft Kinect. Frank and Sandy [18] have used Kinect for interpretation of American Sign Language for 10 different isolated words. Recognition accuracy of 97% was achieved using support vector machine. Yanhua et al. [3] presented recognition system for Japanese sign language using Microsoft Kinect sensor. A method was developed to employ two Kinects for getting more dataset of hand signs for which point cloud library (PCL) was used to get processed data. Zang et al. [6] have used improved SURF algorithm and SVM to recognize static sign using Kinect. Most of these techniques required large number of training samples and are mostly dependent on the signer. In reality, signer independent method is more practical and desirable.

3. METHODOLOGY

Figure 1 gives the overview of proposed system. The system consists of modules such as data acquisition, feature extraction and hand gesture recognition using HMM. In data acquisition module color, depth and skeleton information was captured. In feature extraction with the skeleton joint information, angle was formed and used as a feature and data normalization was done on it. In hand gesture recognition module, maximum likelihood of HMM parameters was derived using the Baum–Welch algorithm for given set of sequences.

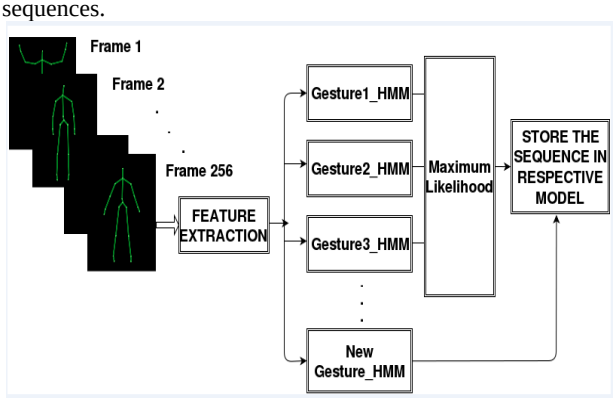

Figure 1 General Flow Diagram of Dynamic HGR

3.1 Dataset

In this paper, total 20 dynamic signs from ISL vocabulary has been considered, which are given in table 1.

Table 1. ISL Gestures signs [14]

Dress-ware			
Shirt	Pant	Tie	Coat
Kurta	Banian	cap	Saree
Other signs			
Sky	Stars	Cloud	Wind
Lightning	sun	Thunder	Sunrise
Sunset	Room	House	Roof

3.2 Skeleton tracking using Kinect camera:

Skeleton joints coordinates was captured using Kinect camera. The coordinate(x, y, z) was the distance of the point along x, y, z axis from the camera. Each frame provides coordinates of all the joints. Here we were basically concerned with three joints- left wrist, right wrist and spine centre. The coordinates of these three joints were tracked from each frame.

3.3 Feature Extraction

Detailed study of ISL vocabulary for mentioned sign shows that we can consider the feature as an angle between two wrists taken from any fixed joint which will not be in motion. Here the angle (θ) was taken from the spine centre joint. As long as the person is not in motion the gesture doesn't involve any movement in the hip. Thus we can define each gesture with a unique angle.

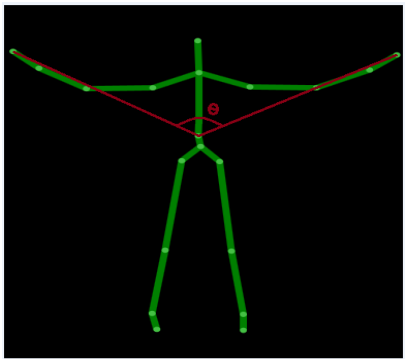

Figure 2 Angle as a feature

3.4 Recognition Using Hidden Markov Model

While analyzing the training set, it was observed that with spine centre as the base of an angle, the maximum angle formed by two wrists ranges from 54° to 162°. Considering this range, the angle was normalized into the range of 1 to 18 values. This normalized feature of each frame form an observation symbol for HMM stage.

Furthermore, 140 frames were found sufficient for conveniently performing any gesture (in above gesture set). Average of angles of 14 frames forming one small gesture was considered. Considering the training set, the hand gesture cannot considerably vary in 14 frames. The combination of 10 states and 18 observation symbols formed a HMM model. The

model is LR model, as the transition of the states is from the left to right without repetition on the same state.

With the observation symbols and hidden states, the trellis diagram for proposed model is given in figure 3.

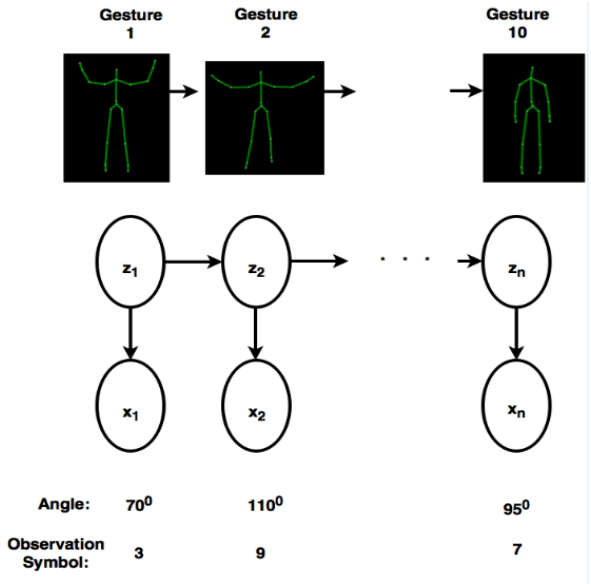

Figure 3 Trellis diagram of proposed system

In above trellis diagram, z_i represents a hidden state and the x_i is observation symbol. There is transition from z_1 to z_2 and so on to z_{10}. z_1 gives the observation symbol x_1, z_2 gives x_2 and so on.

HMM = (π, A, B) where π represents initial vector, A is the transition probability matrix and B refers to emission probability matrix [15].

$$A = \{Z|z_i \rightarrow z_{i+1}\}, where\ i = 1,2,3 \dots 9.$$

$$Z = \{z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8, z_{10}\}$$

$$\pi = \{z_1\} and\ B = \{z_i \rightarrow X\}\ where$$

$$X = \{x_i | i < 11, i > 0\}, where\ x_i = 1,2,3, \dots 18.$$

Gesture made by the hand and their output sign form the hidden state(z) which is not known to the system. Whereas, the feature extracted from each frame (θ) is the observation symbol for HMM.

Based on the previous data collected, we can find the transition probability and the emission probability. Thus HMM equation is:

$$\text{For } P(x_1, x_2, \dots x_n, z_1, z_2, z_3, \dots z_n)$$

$$initial\ state\ is\ P(z_1)\ P(x_1|\ z_1)$$

$$\prod_{k=2}^{n} P(z_k|\ z_{k-1})\ .P(x_k|\ z_k)\ (1)$$

Let us write

$$\boldsymbol{P} = (x_1, x_2, \dots x_n, z_1, z_2, z_3, \dots z_n) as\ P(X, Z)$$

Where,

P (z_1) P ($x_1|z_1$) is the probability of x_1 given z_1. It is the initial state(π). We have added it as it does not have any previous state.

P ($z_k|\ z_{k-1}$) is the probability of z_k given z_{k-1} .This is the transition state. Let us denote it as A.

P ($x_k|\ z_k$) is probability of x_k given z_k .This is the emission state. Let us denote it as Ɛ.

So HMM equation using equation 1 is given in equation 2.

$$P(X, Z) = \Pi(i).Ɛz_1(x_1) \prod_{k=2}^{n} A(z_{k-1}, z_k).Ɛ\ ^{z_k}(x_k) \quad (2)$$

With Hidden Markov Model, following problems are solved.

i) Match most likely system to sequence of observation (will check the familiarity of the unknown gesture with the gestures in the dataset).

ii) Determine hidden sequence generated by sequence of observations (gestures corresponding to the observation symbol- Θ)

In this paper Viterbi algorithm and Baum Welch (BW) algorithm is used for training purpose. In Viterbi algorithm, we find the maximum likelihood of the given sequence to the trained model. Thus the goal is to find Z*=argmaxP(z|x) . Algorithm 3.4.1 and 3.4.2 gives detailed working for feature extraction and recognition respectively.

Algorithm 3.4.1: Feature extraction and processing from tracked skeleton
Procedure FeatureExtractionAndNormalization (Skeleton skel)

Initial: Each gesture is tracking 156 frames. First 13 and last 2 frames are skipped. Frames between 14 to 154 frames are tracked (140 frames). Each gesture is consists of 10 states so in every 14 frames one symbol (state) is derived.

1: frameNo=0; angleSpinSum=0;

2: SymbolArray A[10]=0; //Ten state gesture is initialized to zero initially

3: if (frameNo==14)

4: ALARM; //New Gesture processing will start

5: if(frameNo==154)

6: ALARM; // Ready for next gesture

7: if(frame==155) //154th frame is the last frame

8: Call Procedure HMMRecognition (SymbolArray A)

// Call HMM classification procedure for tracked Skeleton

9: else //extract the angle feature

10: RightWrist = skel. WristRight

// Co-ordinators of right wrist joint

11: LeftWrist = skel.WristLeft

// Co-ordinators of left wrist joint

12: SpineJoint = skel. Spine

// Co-ordinators of spine joint

13: AngleSpine=AngleBetweenTwoVectors(RightWrist – SpineJoint, LeftWrist–SpineJoint)

14: if (frameNo%14 == 0)

// Check number of frame is in multiple of 14

15: Avg_angle = average(AngleSpineSum)

16: Observation_symbol = normalize(Avg_angle)

//Normalize the angle into symbols 1 to 18

17: A = store the Observation_symbol in array A

18: else

19: AngleSpineSum=AngleSpineSum+AngleSpine

End Procedure

Algorithm 3.4.2: Hand gesture recognition using HMM
Procedure HMMRecognition (SymbolArrayA[])

Initial: Array of symbols are transferred into HMM models. Each HMM model represents one gesture. All HMM models of related to one gesture are stored in single file F. So for unknown gesture, search HMM likelihood gesture in parallel.

1: gestureNumber =n;
//n is number of gestures stored in file F

2: noOfSymbols=10; //10 state gesture
3: noOfFiles=20 //20 gesture types

4: for i=1 to noOfFiles
5: SymbolArray sequence [ ] [ ] = SymbolArray [gestureNumber] [noOfSymbols];
// Stored symbols from file F_i
//sequence[1][1→10] to sequence [n][1→10]

6: HiddenMarkovModel hmm (10, 18); // Create instance of HMM having10states and each have 18 possible values.

7: learn = BaumWelchLearning(hmm); //Set Tolerance = 0.0001,Iterations = 10

8: learn.Run(sequences) ; // Train the symbols from the file

9: $likelihood_i$ = hmm.Evaluate(A); // Evaluate likelihood of the array A with respect to the sequences from file

// Check the probability of sequence A with each HMM model

10: end for

11: Gesture= Max($likelihood_i$) // Sequences which contains few errors have higher probability
12: Print Gesture

13: StoreSequence(F_i , A); //Store the sequence into File i.

End Procedure

4. RESULTS

Total 20 ISL signs on four different persons considering scaling variation were considered. Each sign tested on four different persons 10 times and average sign accuracy was calculated. Total average accuracy is calculated by equation (3). Result shows that, on expert signer we have achieved good accuracy and for non-expert signer accuracy decreases. But still, system is designed in such a way that at testing time training set gets updated and helped to increase recognition rate for Non-expert signer.

Sign language recognition is a very challenging problem with limitation and constraints of computer vision and cannot be directly compared with other sign language work due to dependency of various signs, camera and dataset [19].

$$Total\ average\ accuracy(\%) = \sum_{sign=1}^{n} \sum_{person=1}^{m} \frac{correct\ \_recognition}{nm} * 100 \quad (3)$$

Table 2 Accuracy of ISL signs using HMM

Sr.no.	Gesutre	Person 1	Person 2	Person 3	Person 4	Avg. Acc%
1	Shirt	100	100	100	100	100
2	Pant	100	100	100	100	100
3	Tie	90	90	80	90	87.5
4	Coat	100	100	90	100	97.5
5	Kurta	80	100	90	100	92.5
6	Banian	80	80	70	80	77.5
7	cap	90	70	70	90	80
8	Saree	90	80	90	90	87.5
9	Sky	90	70	70	80	77.5
10	Stars	80	80	90	90	85
11	Cloud	90	80	80	90	85
12	Wind	100	100	100	100	100
13	Lightning	100	90	90	100	95
14	Sun	90	80	90	100	90
15	Thunder	100	90	80	100	92.5
16	Sunrise	90	90	90	90	90
17	Sunset	80	80	80	90	82.5
18	House	100	90	80	100	92.5
19	Room	90	90	90	90	90
20	Roof	80	80	80	90	82.5
	Total Avg Accuracy					89.25

5. CONCLUSION AND FUTURE WORK

Dynamic hand gesture recognition is very active research work going in HCI applications. With variety of major applications, sign language recognition is one of the social applications. HMM found most suitable method to deal with dynamic hand gesture recognition from the survey. The aim of this paper was to design and test HMM for real time hand gesture recognition considering single feature such as an angle. So, proposed system tested for two handed 20 ISL dynamic signs on four different people and achieved 89.25% average accuracy. Currently system has been tested on 800 samples however; the algorithm accuracy is directly proportional to the testing examples. This algorithm is suitable for any HCI hand gesture recognition problem where minimum gestures are required. Future work will concentrate to add more feature such as velocity and depth information.

6. REFERENCES

[1] A. Ghotkar, G. Kharate 2015. Dynamic Hand gesture recognition for sign words and Novel Sentence Interpretation Algorithm for Indian Sign Languagne using Microsoft Kinect Sensor. Journal of Pattern Recognition Research, 10(1), pp.25-38.

[2] C. Vogler, D. Metaxas 2001. A Framework for Recognizing the Simultaneous Aspects of American SignLanguage. Computer Vision and Image Understanding pp.358-384

[3] Y. Sun,N. Kuwahara and K. Morimoto 2013. Analysis of recognition system of Japanese sign language using 3D image sensor. IASDR pp.1-7.

[4] K. Stefanovand J.Beskow 2013. A kinect corpus of Swedish sign language Signs. Proceedings Work-shop on Multimodal Corporation pp.1-5.

[5] A. Kuznetsova, L. Leal-Taixe, and B. Rosenhahn 2013. Real-time sign language recognition usingconsumer depth camera pp.83-90.

[6] Z.Hu, L.Yang, L. Luo, Y. Zhang, and X.Zhou 2014.The Research and Application of SURF AlgorithmBased on Feature Point Selection Algorithm. Sensor and Transducers IFSA publishing pp.67-72.

[7] T. Shanablehand K. Assaleh 2007. Arabic sign language recognition in user-independent mode. InProc. int. Conf. Intell. Adv. Syst pp.597-600.

[8] J. Lichenauer, E. Hendriks and M. Reinders 2008 .Sign Language Recognition by Combining Statistical DTW and Independent Classification IEEE Transaction on Pattern Analysis and Machine Intelligence 30(11) pp. 2040-2046.

[9] Elmezain, M. , Al-Hamadi, A. , Michaelis B 2009. Hand trajectory based gesture spotting and recognition using HMM 16th IEEE International Conference on Image Processing (ICIP)

[10] Elmezain, M. , Al-Hamadi, A. , Michaelis, B 2008. A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory Pattern Recognition19th International Conference , ICPR.

[11] Shrivastava R. A 2013. Hidden Markov model based dynamic hand gesture recognition system using OpenCV Advance Computing Conference (IACC).

[12] Gaus, Y.F.A. ,Wong, F 2012. Hidden Markov Model Based Gesture Recognition with Overlapping Hand-Head/Hand-Hand Estimated Using Kalman Filter, Intelligent Systems, Modelling and Simulation (ISMS) Third International Conference.

[13] Moni M.A., Ali A.B.M.S. 2009. HMM based hand gesture recognition: A review on techniques andapproaches, Computer Science and Information Technology ICCSIT 2009.

[14] Instructional Indian sign language video: A project of International human resource development centre (IHRDC) for the disabled. Ramkrishna mission vidyalaya, Coimbatore.http://indiansignlanguage.org.

[15] Wilson, A.D. Media Lab. MIT, Cambridge, Bobick, A.F. Parametric hidden Markov models for gesture recognition, Pattern Analysis and Machine Intelligence, IEEE Transactions ,21(9).

[16] P. Kishore, Rajesh Kumar, E. Kiran Kumar and S.Kishore 2011. Video Audio Interface for Recognizing Gestures of Indian Sign Language. International Journal of Image Processing (IJIP),5(4), pp. 479-503.

[17] J. Han, L. Shao, D. Xu and J.Shotton 2013. Enhanced Computer Vision with Microsoft Kinect Sensor: A Review. IEEE transaction on Cybernetics,43(5), pp.1318-1334.

[18] F. Huang and S. Huang 2011. Interpreting American Sign Language with Kinect pp.1-5.

[19] G. Kharate and A. Ghotkar 2016. Vision based Multi-feature Hand gesture Recognition for Indian Sign Language Manual Signs. International Journal on Smart Sensing and Intilligent System,9(1),pp. 124-147.