Kinect Based Real-time Gesture Spotting Using HCRF

Mahesh Chikkanna
National Institute of Technology Karnataka, Surathkal
Mangalore, India 575 025
Email: cmaheshster@gmail.com

Ram Mohana Reddy Guddeti
National Institute of Technology Karnataka, Surathkal
Mangalore, India 575 025
Email:profgrmreddy@gmail.com

***Abstract*—The sign language is an effective way of communication for deaf and dumb people. This paper proposes, developing the gesture spotting algorithm for Indian Sign Language that acquires sensory information from Microsoft Kinect Sensor. Our framework consists of three main stages: hand tracking, feature extraction and classification. In the first stage, hand tracking is carried out using frames of Kinect. In second stage, the features of Cartesian system (velocity, angle, location) and hand with respect to body are extracted. K-means is used for extracting the codewords of features for HCRF. In the third stage, Hidden Conditional Random Field is used for classification. The experimental results show that HCRF algorithm gives 95.20% recognition rate for the test data. In real-time, the recognition rate achieves 93.20% recognition rate.**

***Keywords*—Hidden Conditional Random Field, Kinect, Gesture Spotting, Indian Sign Language**

I. Introduction

Indian Sign Language (ISL) is derived from British Sign Language (BSL) and French Sign Language (FSL) for communication of deaf and dumb people. Few research works have been carried out in ISL recognition and interpretation using image processing/vision techniques. But these work were involving hand postures only and not much work is carried in the gesture involving hand moments in recognition of ISL. The method for recognizing the static gestures is called gesture recognition whereas recognition of the gestures involving motions of the hand is called as gesture spotting as shown in Fig 1.

Fig. 1: Gesture Spotting

For the gesture spotting, first the hand detection from the background should be done for tracking it. After detecting the hand, points of the gesture path are obtained from it. The feature extraction is carried out on the obtained gesture path. The extracted feature is given to the classification model for recognition of the gesture.

Nasser H. Dardas and Emil M. Petriu [1], Nasser H. Dardas and Nicolas D Georgan [2] proposed the hand detection using face subtraction, skin detection and hand postures contours comparison algorithm. The detected face will be subtracted by replacing face area with a black circle. After subtracting the face, detection of the hand is done using the HSV color model of skin tone. The HSV color model of hand detection fails to detect the hand correctly in the clustered background were other parts of the body are not covered and it is computational expensive. In the feature extraction Q. Yuan, S. Sclaroff, and V. Athitsos [3] used the block-based matching method for extracting the feature. Daniel Kelly et al [4] proposed the feature extraction carried out by finding the left hand position, right hand position, direction of the motion of the hands , hand position relative to each other. Hee-Deok Yang et. al [5] proposed the different set of features to be extracted with respect to current position of the hand in the gesture path. The classification is important and final step of the model. For training the system different classification algorithm are used. Dynamic time wrapping algorithm for the training and spotting the gesture (sign) pattern in the given values of hand movements was proposed by Alon Jonathan [6]. In Dynamic time wrapping the recognition time increases, with increases in the number of gestures. In recognition of the American Sign Language for fun and play Z. Zafrulla et. al [7] proposed a Hidden Markov Model as the classifier. M. Elmezain [8], [9], [10], [11] proposed the use of Hidden Markov Model has the classifier for the gesture spotting approach for the recognition of the arabic alphabets and numbers (0-9). In [12][13] author propose the use of Hidden Conditional Random Field for the gesture spotting. In [12] the Hidden Conditional Random Field performs better than Hidden Markov Model in the recognition of the same gestures.

The rest of the paper is organized as follows: Hand tracking in Section II, Feature extraction in Section III, Classification in Section IV, Expermiental Results in Section V, Conclusion and Future Work in Section VI.

II. Hand Tracking

The Microsoft Kinect is used for the purpose of tracking the hand. Different types of frames are provided by Microsoft Kinect. Different frames are skeleton frame, color frame and depth frame. The skeleton which is near to the Kinect is called as Primary Skeleton. Since the gesture done by the person who is nearer to the Kinect sensor has to be recognized. The

978-1-4673-6217-7/13/$31.00 ©2013 IEEE

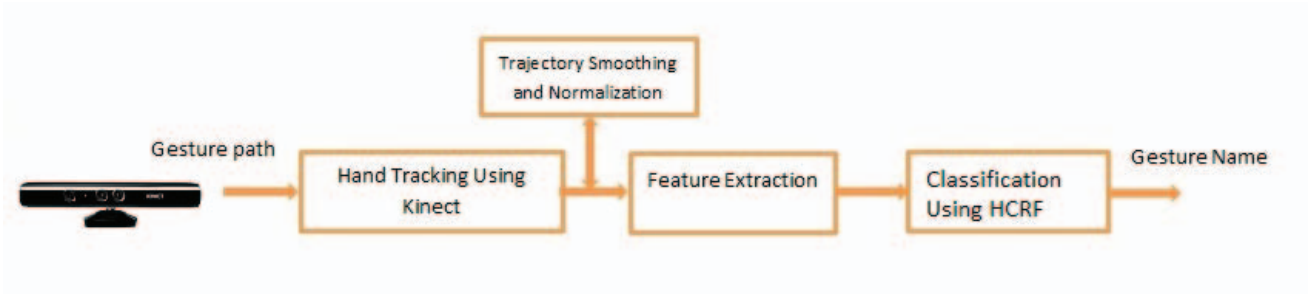

Fig. 2: Gesuture Spotting FrameWork Using HCRF

Algorithm 1 assumes that the length of the hand is double the distance between wrist and hand point in the color image. So, the hand point is taken has centre point(x_c, y_c) and wrist point (x_1, y_1)as one of the point using which the another point (x_2, y_2) is calcuated using mid point formula of Equation 1 .The hand tracking is as shown in Algorithm 1.

$$(x_c, y_c) = (\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}) \quad (1)$$

Algorithm 1 Hand Tracking Algorithm

▷ %Input: skeletonframe: the skelton frame of the kinect% ▷ %Ouput: points for covering the hand in color frame%

$pskeleton = getPrimaryFrame(skeletonframe)$
if $pskeleton\ is\ stable$ **then**
$wPoint = getJointPoint(pskeleton, wristJoint)$
$hPoint = getJointPoint(pskeleton, handJoint)$

▷ % mapping skeleton points to color frame%
$cWristPoint = MapPoint(wPoint)$
$cHandPoint = MapPoint(hPoint)$

▷ %distance between wrist and hand point%
$length = distance(cWristPoint, cHandPoint)$

$points$ = Using circle of radius $length$ with $cHandPoint$ as center obtain the upper and lower points of rectangle outside
return $points$
end if

The final result of the hand detection after segmentation using the Algorithm 1 is shown in Fig 3.

III. Feature Extraction

Selecting a good set of features is very important for the spotting of the gestures. There are three basic features: location, orientation and velocity. The first location feature is L_c which measures the distance from the centroid point to all points of gesture path. The second location feature is L_{sc} which is computed from the start point to the current point of gesture path using the Equation. 2.

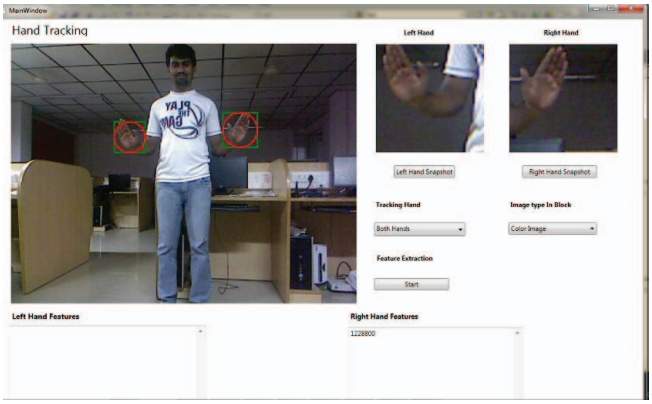

Fig. 3: Hand Image After Segmentation

The second basic feature is the orientation which gives the direction of the hand when traverses in space during the gesture making process. Orientation feature according to the centroid of gesture path (θ_{1t}), the orientation between two consecutive points $(\theta_{2t}$) and the orientation between start and current gesture point $(\theta_{3t}$) is calcuated using the Equation. 3

The third basic feature is velocity (V_t) which plays an important role during gesture recognition phase. The velocity is calcuated using the Equation. 5.

$$length = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2)$$

$$\theta = \tan^{-1}(\frac{y_2 - y_1}{x_2 - x_1}) \quad (3)$$

$$(C_x, C_y) = \frac{1}{n}(\sum_{t=1}^{N} x_t, \sum_{t=1}^{N} y_t) \quad (4)$$

$$V_t = \sqrt{(\frac{x_t - x_{t-1}}{t})^2 + (\frac{y_t - y_{t-1}}{t})^2} \quad (5)$$

Another feature along with the above 6 features is added into the list of features to increase the recognition accuracy. The body is divided into 6 parts and depending on the position of the hand H_v as shown in Fig. 4 at a particular moment in time.

A. Trajectory Smoothing

The hand gesture path is determined by tracking hand using kinect as described in the Section II. The input points of hand from kinect are usually unstable. So, it will cause frequent, sharp changes in the points. In order to efficiently overcome these unexpected changes, the trajectory points are smoothed. The previous hand point (x_{t-1}, y_{t-1}), current point (x_t, y_t) and next point (x_{t+1}, y_{t+1}) using Equation 6. In the Fig 5 we can see the different between the gesture path before and after trajectory smoothing.

$$(x_s, y_s) = (\frac{x_{t-1} + x_t + x_{t+1}}{3}, \frac{y_{t-1} + y_t + y_{t+1}}{3}) \quad (6)$$

B. Normalization

The extracted features are normalized or quantized to obtain the discrete symbols which are used as an input to HCRF. The basic features such as location and velocity are normalized with different scalar values ranging from $scale_{min}$ to $scale_{max}$ when used separately. The scalar values increase the robustness for selecting the normalized feature values. The normalization is done as follows:

1) Find the maximum value of N values in the current gesture path using Equation 7.

$$Norm_{max} = max_{i=1}^{N}(Norm_i) \quad (7)$$

2) Scaling the values between $scale_{min}$ to $scale_{max}$ using the Equation 8.

$$Norm_i = \frac{Norm_i}{Norm_{max}}.Scal \quad (8)$$

where $Norm_i$ represents the feature vector of dimension i to be normalized and $Norm_{max}$ is the maximum value of the feature vector which is determined from all the N points in the gesture trajectory.

The normalization of the orientation features is estimated by dividing them by 10^0, 20^0, 30^0 or 40^0 to obtain their code words which are employed for HCRF as shown in Fig. 6 for dividing by 45^0. K-means clustering algorithm is used to classify the gesture feature into K clusters on the feature space.

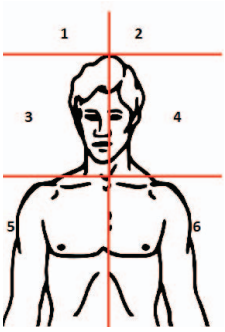

Fig. 4: Feature values after dividing the body into 6 part for hand position as a feature.

Fig. 5: Before and After Trajectory Smoothing of Gesture Path

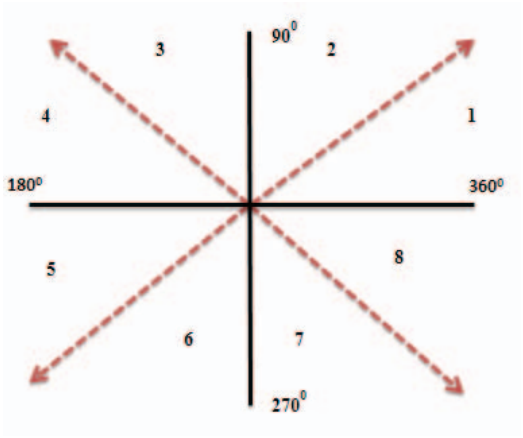

Fig. 6: The direction code from 1 to 8 in case of dividing the orientation by 45^0

IV. Classification

Hidden Conditional Random Fields is used for the classification of the gesture. It is used for training the system and recognition the gestures. From HCRF we learn mapping of observations x to class labels $y \in Y$ where x is a vector of m local observations, $x = \{x_1, x_2, ...x_m\}$ and each local observation x_j is represented by a feature vector $\varphi(x_j) \in R^d$.

An HCRF models the conditional probability of a classlabel given a set of observations by Equation 9.

$$P(y|x, \theta) = \Sigma_s P(y, s|\theta) = \frac{\Sigma_s e^{\Phi(y,s,x;\theta)}}{\Sigma_{y' \in Y, s \in S^m} e^{\Phi(y',s,x;\theta)}} \quad (9)$$

where $S = \{s_1, s_2, s_3...s_m\}$ each $s_i \in S$ captures certain underlying structure of each class and S is the set of hidden states in the model. If we assume that s is observed and that there is a single class label y then the conditional probability of s given x becomes a regular CRF. The potential function $\Phi(y, s, x; \theta) \in R$, parameterized by θ, measures the compatibility between a label, a set of observations and a configuration of the hidden states. For $\Phi(y, s, x; \theta)$ function is defined as the potential function parameterized by θ is as shown below:

$$\Phi(y, s, x; \theta) = \Sigma_{t=1}^{T} \lambda_{ij} f_{ij}(y_t, y_{t-1}, x, s) + \Sigma_{t=1}^{T} \mu_{ij} f_{ij}(y_t, x, s) \quad (10)$$

We use the following objective function in training the parameters:

$$L(\theta) = \Sigma_{i=1}^{n} \log P(y|x, s, \theta) \quad (11)$$

where n is the number of training sequences.We use gradient ascent to search for the optimal parameter values, $\theta^* = arg\max_\theta L(\theta)$. For our experiments we used a Resilent propogation optimization technique.

V. Experimental Results

Our method introduces a method to recognize the Indian Sign Language involving the motion of single hand using HCRF. The method used C# for implementation. In our experiment 650 gestures are recorded using the Kinect gesture training system of ten Indian Sign Language gestures. Each gesture of 65 training data are recorded, out of which 40 gestures of each are used for training and the rest are used for testing.

We test the gesture spotting by using HCRF algorithm. For the test data gesture spotting as shown in using gives 95.20% accuracy as shown in Table I. For Real-time gesture spotting using HCRF the result as shown in Table II it gives 93.20% accuracy in recognition of the gesture in real-time.

TABLE I: Gesture Spotting using HCRF

Gesture Name	Test Data	Recognition	Recognition%
Elephant	25	25	100.00%
Wish	25	22	88.00%
Then	25	22	88.00%
Now	25	21	84.00%
Pen	25	24	96.00%
Internal	25	25	100.00%
Output	25	25	100.00%
Aeroplane	25	25	100.00%
Until	25	25	100.00%
Tomorrow	25	24	96.00%
Recognition%	250	238	95.20%

TABLE II: Real Time Gesture Spotting using HCRF

Gesture Name	Test Data	Recognition	Recognition%
Elephant	25	24	96.00%
Wish	25	23	92.00%
Then	25	23	92.00%
Now	25	25	100.00%
Pen	25	20	80.00%
Internal	25	23	92.00%
Output	25	21	84.00%
Aeroplane	25	25	100.00%
Until	25	25	100.00%
Tomorrow	25	24	96.00%
Recognition%	250	233	93.20%

VI. Conclusion and Future Work

This paper proposes a framework to recognize Indian Sign language involving single hand motion from Kinect using HCRF. This framework combines the features of location, orientation, velocity, position of with respect to body as shown in Fig 4. The database contains 650 kinect based gestures recorded by us. The framework yield recognition rate of 95.20% for of test data and 93.20% in real-time. The future research will be focused on recognizing the motion involving both the hands.

References

[1] N. Dardas and E. Petriu, "Hand gesture detection and recognition using principal component analysis," in *Computational Intelligence for Measurement Systems and Applications (CIMSA), 2011 IEEE International Conference on*, Sept., pp. 1–6.

[2] N. Dardas and N. Georganas, "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques," *Instrumentation and Measurement, IEEE Transactions on*, vol. 60, no. 11, pp. 3592–3607, Nov.

[3] Q. Yuan, S. Sclaroff, and V. Athitsos, "Automatic 2d hand tracking in video sequences," in *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on*, vol. 1, Jan., pp. 250–256.

[4] H.-K. Lee and J. Kim, "An hmm-based threshold model approach for gesture recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 10, pp. 961–973, Oct.

[5] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 7, pp. 1264–1277, July.

[6] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 9, pp. 1685–1699, Sept.

[7] Z. Zafrulla, H. Brashear, P. Presti, H. Hamilton, and T. Starner, "Copycat: An american sign language game for deaf children," in *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, March, pp. 647–647.

[8] M. Elmezain, A. Al-Hamadi, and B. Michaelis, "Hand trajectory-based gesture spotting and recognition using hmm," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, Nov., pp. 3577–3580.

[9] M. Elmezain and A. Al-Hamadi, "Gesture recognition for alphabets from hand motion trajectory using hidden markov models," in *Signal Processing and Information Technology, 2007 IEEE International Symposium on*, Dec., pp. 1192–1197.

[10] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, "A hidden markov model-based continuous gesture recognition system for hand motion trajectory," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, Dec., pp. 1–4.

[11] M. Elmezain, A. Al-Hamadi, and B. Michaelis, "A robust method for hand gesture segmentation and recognition using forward spotting scheme in conditional random fields," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, Aug., pp. 3850–3853.

[12] S. B. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 1521–1527.

[13] Y. Song, D. Demirdjian, and R. Davis, "Multi-signal gesture recognition using temporal smoothing hidden conditional random fields," in *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 388–393.