2014 Fourth International Conference on Communication Systems and Network Technologies

A Conditional Random Field based Indian Sign Language Recognition System under Complex Background

Ananya Choudhury*, Anjan Kumar Talukdar* and Kandarpa Kumar Sarma*
*Dept. of Electronics and Communication Engineering
Gauhati University, Guwahati-14, Assam, India
Email: achoudhury50@gmail.com, anjan.nov@gmail.com, kandarpaks@gmail.com

***Abstract*—Sign languages are natural languages that use different means of expression for communication in everyday life. Automatic sign language recognition has a significant impact on human society as it can provide an opportunity for the deaf to communicate with non-signing people without the need for an interpreter. In this paper, we present a Conditional Random Field (CRF) based Indian Sign Language (ISL) recognition system which is effective under complex background using a novel set of features. Hand segmentation is the most crucial step in every hand gesture recognition system since if we get better segmented output, better recognition rates can be achieved. The proposed system also includes efficient and robust hand segmentation and tracking algorithm to achieve better recognition rates.**

Keywords*—Sign Language Recognition, Complex background, Skin Color Segmentation, Frame Differencing, Contour Processing, Conditional Random Field.

I. Introduction

Gesture is basically a movement of the body part/parts which contain information or feelings. Gesture recognition is a mechanism through which a system can understand the meaning of any gesture. Hand Gesture can be subdivided into two types, firstly global motion where the entire hand moves and secondly local motion (or posture) where only the fingers move [1]. In sign language, both local and global motions are considered.

Automatic sign language recognition has attracted researchers for long because it can provide an opportunity for the deaf to communicate with non-signing people without the need for an interpreter. Most of the research work as in [2], [3], [4], [5] have concentrated mainly on recognition of single-handed gestures in complex background. Recognition of double-handed gestures in complex background and background containing multiple gesturers has not been taken into consideration. So, the aim of this paper is to report an efficient ISL recognition system which can recognize both single-handed and double-handed signs under complex background using a novel set of features. Problems such as complex background, background involving multiple gesturers and multiple skin coloured image regions are effectively handled by our system.

This paper is organized as follows: basic theoretical considerations are presented in Section II. Section III describes the proposed model. Section IV contains the experimental results. Finally, Section V concludes the work.

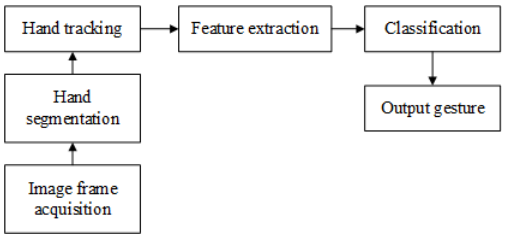


Fig. 1: Block diagram of a basic hand gesture recognition system

II. Basic Theoretical Considerations

The generalized block diagram for hand gesture recognition system is shown in Fig. 1. Firstly, the image frame is acquired from the input video. The next step is segmentation which partitions an image into its constituent parts or objects. Hand tracking is a high-resolution technique that is employed to know the consecutive position of the hands of the user. After the successful tracking there is a need to extract the important feature points from the available data points of the track path. In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction [6]. After feature extraction, classifier plays a vital role in gesture recognition process. Classifier is a statistical method that takes feature set as input and gives a class labeled output, which are required output gestures [7]. The classifiers which are widely used in gesture recognition are HMM (Hidden Markov Model) and CRF (Conditional Random Field) model.

A. Hand Segmentation

In gesture recognition, hand segmentation is the pre-requisite to track the movement of the hand.

1) Hand segmentation techniques: The various types of commonly used hand segmentation techniques are described below:

- Skin Color Segmentation: In this model, the input image is first taken and then the image in RGB color

978-1-4799-3070-8/14 $31.00 © 2014 IEEE
DOI 10.1109/CSNT.2014.185

space is converted into HSI color space as in this color space, Hue(H) and Saturation(S) are independent of illumination and reflectance. Then thresholding is done to convert the HSI image into a binary image. Noise is minimized using morphological operations like erosion, dilation etc. Disadvantage of this system is that if the background has any object having the same color as the hand, noise will be very high [8].

- Frame Differencing: This is a very simple background subtraction method where one frame is subtracted from another and then any difference that is big enough is labeled as foreground. This process tends to catch the edges of moving objects which in our case would be the body parts. Disadvantage of this system is that if the lighting conditions change abruptly then there is a change in pixel value where the light intensity changed and additive noise contributes to the output [9] [10].

2) Contour Matching: Contours are sequences of points that represent a line/curve in an image. Every entry in the sequence encodes information about the location of the next point on the curve/line. Contour matching is most often used for recognizing and classifying image objects [11].

- Contour matching algorithm: The contour model, containing pixels that are on the edge of (a part of) the object is placed on all possible positions in the search image (or another contour), computing a match value for every position. The match value is now based on the edge pixels in the contour model only. All the search image/contour pixels that correspond to an edge pixel in the contour model are added. The higher this value is, the better the resemblance between the contour model and search image/contour [12]. Contour matching proves to be a more preferable choice over template matching as the matching algorithm is restricted to only the edge pixels rather than the whole image as in template matching. Hence it faster, yields sharp matches and invariant under imaging transformations like scaling, translation, rotation, intensity. However, the hindrance of this method is that since a smaller number of pixels is involved in recognition, the influence of each pixel is greater, i.e. a few deviating pixels (due to noise, distortion) may hinder recognition. This has to be taken care of [13].

B. Hand Tracking

In gesture recognition, hand tracking is done to find out the hand trajectories made during gesticulation. One of the most commonly used method for hand tracking is calculation of the centroids of the segmented hand and subsequently connecting them to get the gesture trajectories.

1) Calculation of centroids: An image moment is a particular weighted average (moment) of the image pixels' intensities, or a function of such moments. Image moments are useful to describe objects after segmentation. Simple properties of the image which are found via image moments include area (or total intensity), its centroid, and information about its orientation. For a 2D binary image or contour I(x,y), an image moment M_{ij} is calculated using [11]:

$$M_{ij} = \sum_{i=1}^{n} I(x,y) x^i y^j \tag{1}$$

where i is the x-order and j is the y-order. The summation is over all the pixels of the contour boundary or image (as denoted by n in the equation). The centroid (x,y) of the binary image or contour is given by:

$$(x,y) = (\frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}}) \tag{2}$$

where M_{00} i.e. the zeroth moment gives the number of pixels (area) of the binary image or contour.

C. Feature Extraction

Transforming the input data into a set of features is called feature extraction. Feature extraction is important in terms of giving input to a classifier by means of which it can understand the meaning of a gesture. The major features for gesture recognition are templates, global transformations, zones and geometric features [14].

D. CRF Framework For Recognition

CRFs are a framework based on conditional probability approaches for segmenting and labeling sequential data. CRFs use a single exponential distribution to model all labels of given observations. In CRFs, the probability of label sequence *Y*, given observation sequence *X*, is found using a normalized product of potential functions. Thus the conditional probability is given by [15]:

$$P_\theta(Y|X) = \frac{1}{Z_\theta(X)} exp(\sum_{i=1}^{n} F_\theta(Y_{i-1}, Y_i, X, i)) \tag{3}$$

In equation (3),

$$F_\theta(Y_{i-1}, Y_i, X, i) = \sum_v \lambda_v t_v(Y_{i-1}, Y_i, X, i) + \sum_m \mu_m s_m(Y_i, X, i) \tag{4}$$

where,
$t_v(Y_{i-1}, Y_i, X, i)$ is a transition feature function of observation sequence *X* at positions *i* and *i-1*. A transition feature function indicates whether a feature value is observed between two states or not.
$s_m(Y_i, X, i)$ is a state feature function of observation sequence at position *i*. A state feature function indicates whether a feature value is observed at a particular label or not.
Y_{i-1} and Y_i are labels of observation sequence *X* at position *i* and *i-1*.
n is the length of the observation sequence.
λ_v and μ_m are weights of transition and state feature functions, respectively.
$Z_\theta(X)$ is the normalization factor.

$$Z_\theta(X) = \sum_Y exp(\sum_{i=1}^{n} F_\theta(Y_{i-1}, Y_i, X, i)) \tag{5}$$

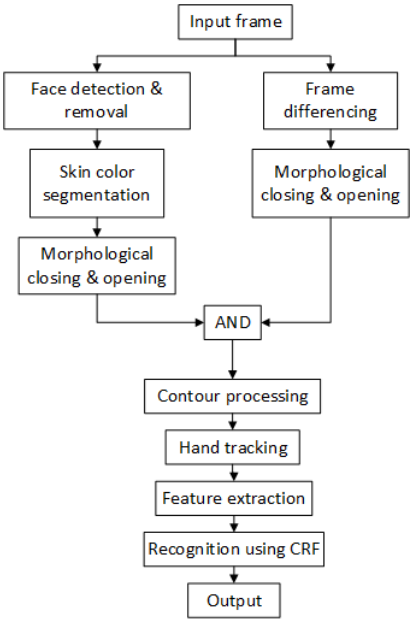


Fig. 2: Block diagram of proposed model

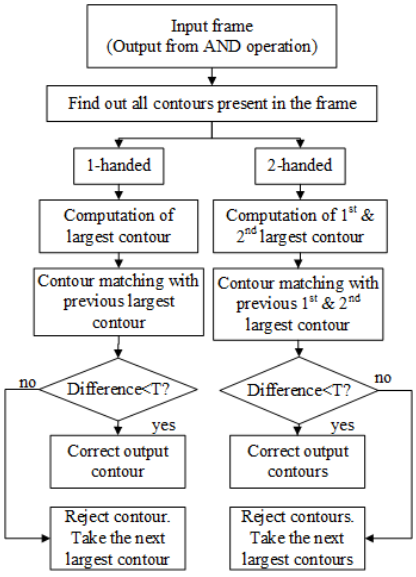


Fig. 3: Flowchart of contour processing stage

III. Proposed Model

The block diagram of the proposed model is shown in Fig. 2. The detailed description of all the steps undergone is illustrated as follows:

A. Hand Segmentation

At first, the input frames are captured from webcam and face detection and removal using Haar Classifier is done to mask out the face region. After that skin color segmentation is done to segment out the hand region. This is followed by some morphological operation to filter out noise and to fill up holes (output O1). Simultaneously, on the other side motion detection is carried out by frame differencing of the input frames. This is again followed by some morphological operation (output O2). In the next step logical AND operation is performed between outputs O1 and O2 to get the final hand segmented output. The output after AND operation is effective only in case of simple and complex background, but in case of background having multiple gesturers the last step i.e. contour processing is done irrespective of one-handed and two-handed gesture to get the correct gesture output.

The contour processing stage is shown in Fig. 3. In this stage, at first all the contours present in the input frame are found out. Followed by it, the largest contour having largest area is computed for one-handed gesture. In case of two-handed gesture, the first largest and second largest contour is determined.There might be a situation where an unintended gesturer may have contour greater than that of the actual gesturer. In such situations, an incorrect gesture might be reported as a correct gesture output. So, after computation of the largest contour, contour matching is done by finding the difference with the previous largest contour in case of one-handed gesture and for two-handed gesture, it is performed with the previous first largest and second largest contour. For contour matching, a training and testing method is employed. A threshold value of difference (T) is selected empirically and if the obtained value of difference is less than T, it will be interpreted as correct gesture otherwise it will be rejected.

B. Hand Tracking

The hand tracking stage is shown in Fig. 4. In this stage, the centroid of the contour obtained from contour processing step is determined using moments calculation and subsequently used for getting the hand trajectory. In case of one-handed gestures, the centroid of the largest contour in the current frame is found out and it is connected to the centroid of the largest contour in the previous frame. In case of two-handed gestures, the criteria used for classifying a detected contour to the left-hand contour(or right-hand contour) is that the distance between the centroids of the left-hand contour (or right-hand contour) will be less than that of the distance between the centroids of left-hand and right-hand contour. So, given the first largest and second largest contour in the previous frame, the best match in the current frame is determined by finding the distance between the centroids of first largest contour in the previous frame and in current frame, and second largest contour in previous and in current frame.

Let, prevC1 be the centroid of 1st largest contour in previous frame and currC1 be the centroid of 1st largest contour in current frame.

prevC2 be the centroid of 2nd largest contour in previous frame and currC2 be the centroid of 2nd largest contour in current

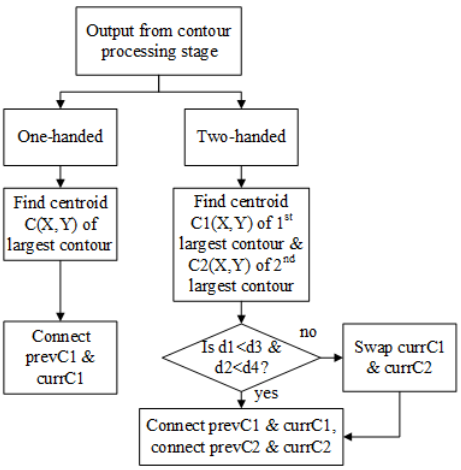


Fig. 4: Flowchart of hand tracking stage

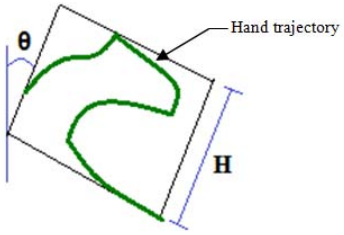


Fig. 5: Extraction of height (H) and orientation (θ) features

frame
d1 be the distance between prevC1 and currC1
d2 be the distance between prevC2 and currC2
d3 be the distance between prevC1 and currC2
d4 be the distance between prevC2 and currC1
Then the proposed algorithm states that-
If d1<d3 and d2<d4, the centroids currC1 and currC2 will remain unchanged.
Else centroids currC1 and currC2 are interchanged.

C. Feature Extraction

Given the hand trajectory obtained from hand tracking stage, it is approximated by a minimum area bounding rectangle. The height (H) and orientation of the rectangle with respect to the vertical (θ) is obtained as shown in Fig. 5. Thus, the features extracted for one-handed and two-handed signs, respectively are shown in Table I.

D. Recognition

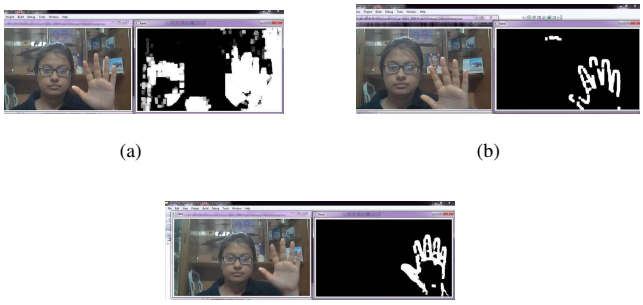
A CRF model for one-handed signs and two-handed signs respectively, is separately trained using the features shown in Table I. To measure the accuracy of the proposed model, the sign spotting/recognition rate is calculated by-

$$R = \frac{C}{N} X 100\% \quad (6)$$

where C is the number of correct spotting and N is the number of test signs.

TABLE I: Features extracted for one-handed and two-handed signs

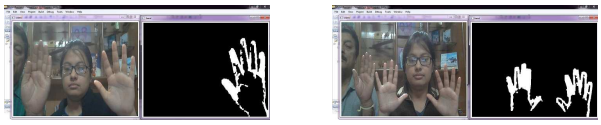
Signs	Features	Meaning
One-Handed	H	Height of minimum area rectangle
	θ	Orientation of minimum area rectangle w.r.t vertical
Two-Handed	H_L	Height of left minimum area rectangle
	H_R	Height of right minimum area rectangle
	θ_L	Orientation of left minimum area rectangle w.r.t vertical
	θ_R	Orientation of right minimum area rectangle w.r.t vertical



(a) (b)

(c)

Fig. 6: Segmented output under complex background using: (a) skin color segmentation (b) frame differencing (c) combination of skin color segmentation and frame differencing



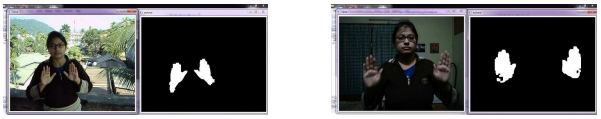
(a) (b)

Fig. 7: Segmented output using proposed method for a complex background having multiple gesturers: (a) one-handed gesture (b) two-handed gesture

IV. Results

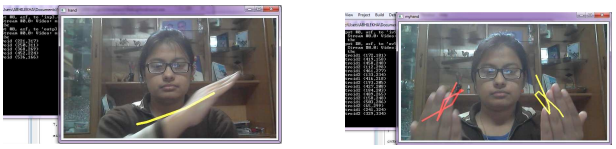
The results obtained from experiments are shown in Fig. 6, 7, 8 and 9 respectively.

Fig. 6(a), Fig. 6(b) and Fig. 6(c) shows the inputs and corresponding outputs obtained using skin color segmentation, frame differencing and combination of the above two methods respectively for complex background. Fig. 7(a) and Fig. 7(b) shows the inputs and outputs of the proposed model under complex background with multiple gesturers for one-handed and two-handed gesture respectively. The proposed model has also been tested under day-light and dim-light conditions. Fig.



(a) (b)

Fig. 8: Segmented output using proposed model for a two-hand gesture input under: (a) day-light condition (b) dim-light condition



(a) (b)

Fig. 9: Hand trajectory for: (a) one-handed sign 'aeroplane' (b) two-handed sign 'study'

8(a) and Fig. 8(b) show the results respectively. The hand tracking output for a one-handed sign and two-handed sign is shown respectively in Fig. 9(a) and Fig. 9(b).
The recognition results obtained from the trained CRF model for one-handed and two-handed case are shown in Table II and Table III respectively. The number of train samples per class and the number of test samples per class were taken 10 each.

V. Conclusion

Hence we have developed an efficient ISL recognition system for spotting of both single handed and double handed signs under complex background using a novel set of features. Moreover, our system doesnot break down even in the presence of multiple signers in the background. The proposed system has also been tested under different lighting conditions. The hand segmentation and tracking results show the effectiveness of the system. Experiments have demonstrated that our system could detect one-handed ISL signs with a 90.0% recognition rate and two-handed ISL signs with a 86.0% recognition rate. Near term future work will include extending the proposed model to recognize continuous ISL signs and also to improve the recognition rates.

TABLE II: Recognition results with one-handed ISL signs

Signs	N	C	R(%)
Aeroplane	10	10	100
Bucket	10	7	70
Round	10	8	80
East	10	10	100
West	10	10	100
Overall Recognition Rate			90

TABLE III: Recognition results with two-handed ISL signs

Signs	N	C	R(%)
Study	10	10	100
Triangle	10	9	90
Model	10	9	90
Coat	10	7	70
Bird	10	8	80
Overall Recognition Rate			86

Acknowledgment

We would like to offer our heartfelt gratitude to all the teachers and fellow students for their help and suggestions in completion of this paper. The authors are also thankful to the reviewers for their constructive review reports.

References

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

[2] Y. Hamada, N. Shimada and Y. Shirai, "Hand Shape Estimation under Complex Backgrounds for Sign Language Recognition" , Proceedings of 6th Intetnational Conference on Automatic Face and Gesture Recognition, pp. 589-594, May 2004.

[3] Q. Zhang, F. Chen and X. Liu, "Hand Gesture Detection and Segmentation Based on Difference Background Image with Complex Background", Proceedings of the 2008 International Conference on Embedded Software and Systems, Sichuan, 29-31 July 2008, pp. 338-343.

[4] J.S. Lee, Y.J. Lee, E.H. Lee and S.H. Hong, "Hand Region Extraction and Gesture Recognition from Video Stream with Complex Background through Entropy Analysis," Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1513-1516, San Francisco, September 2004.

[5] J. R. Pansare, H. Dhumal, S. Babar, K. Sonawale and A. Sarode, "Real Time Static Hand Gesture Recognition System in Complex Background that uses Number system of Indian Sign Language", International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), vol. 2, no. 3, pp. 1086-1090, March 2013.

[6] R.O. Duda, P.E. Hart and D.G. Stork : Pattern Classification, 2^{nd} Ed., Wiley India, Indian Reprint, New Delhi, 2009.

[7] S. Theodoridis and K. Koutroumbas : Pattern Recognition, 3^{rd} Ed., Elsevier, USA, 2006.

[8] S.L. Phung, A. Bouzerdoum and D. Chai, "Skin Segmentation Using Color Pixel Classification: Analysis and Comparison", IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 27, no. 1, pp 148-151, 2005.

[9] C. Manresa, J. Varona, R. Mas and F.J. Perales, "Real-Time Hand Tracking and Gesture Recognition for Human-Computer Interaction", Computer Vision Center Universitat Autonoma de Barcelona, Barcelona, pp-1-3, Spain,2000.

[10] Y.C. Lu, "Background Substraction Based Segmentation Using Object Motion Feedback", First International Conference on Robot Vision and Signal Processing(RVSP), pp 224-227, Kaohsiung ,2011.

[11] G. Bradski and A. Kaehler, Learning OpenCV, 1^{st} Ed., O'Reilly Media, USA, September 2008.

[12] P.J.H.M. Boots and D.Van Schenk Brill, "Object Recognition by Contour Matching", Fontys University of Professional Education, IPA Research Centre Eindhoven, Netherlands.

[13] D. Zhang and G. Lu, "Review of Shape Representation and Description Techniques", Pattern Recognition, vol. 37, no. 1, pp. 1-19, 2004.

[14] M.H. Yang, N. Ahuja and M. Tabb, "Extraction of 2D Motion Trajectories and its Application to Hand Gesture Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 8, pp. 1061-1074, 2002.

[15] H.D. Yang, S. Sclaroff and S.W. Lee, "Sign Language Spotting with a Threshold Model Based on Conditional Random Fields", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 7, pp. 1264-1277, July, 2009.