International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 4, July-Aug-2020, ISSN (Online): 2395-566X

Indian Sign Language Recognition

Mohit Patil, Pranay Pathole, Hrishikesh Patil, Ashutosh Raut, Prof. S S Jadhav
Dept. of Computer Science Engg.
AISSMS College of Engineering, Pune
Pune, MH,India

Abstract:- This paper introduces a system that can recognize poses and hand gestures of Indian Sign Language in real time using grid-based features. This system tries to reduce the communication gap between listening and speaking disabled and the rest of society. Existing solutions either provide relatively low precision or do not work in real time. The system provides good results in both parameters. Sign Language is captured from a smartphone camera and its frames are sent to a remote server for further processing. Techniques such as Face detection, object stabilization and skin color are used. Segmentation is also used for hand detection and tracking. The image undergoes a gridbased function extraction technique representing the placement of the hand in the form of Function Vector, Hand postures are classified using the k-Nearest Neighbors algorithm. However, by gesture classification, movement and intermediate positions of the hand observation sequences are entered into the chains of the hidden Markov model corresponding to the pre-selected gestures defined in Indian Sign Language.

***Keywords*- Computer Vision, OpenCV, Segmentation.**

I. INTRODUCTION

Indian Sign Language (ISL) is a sign language used by hearing and speech impaired people to communicate with other people. The research presented in this paper pertains to ISL as defined in the Talking Hands website. ISL uses gestures for representing complex words and sentences. It contains 28 hand poses including 8 digits, and 20 letters. Amongst the letters in ISL, the letters 'h', 'j' are represented by gestures and the letter 'v' is similar to digit 2. The system is trained with the hand poses in ISL. Most people find it difficult to comprehend ISL gestures. This has created a communication gap between people who understand ISL and those who do not.

One cannot always find an interpreter to translate these gestures when needed. To facilitate this communication, a potential solution was implemented which would translate hand poses and gestures from ISL in realtime. It comprises of an Android smartphone camera to capture hand poses and gestures, and a server to process the frames received from the smartphone camera. The purpose of the system is to implement a fast and accurate recognition technique.

The system described in this paper successfully classifies all the 33 hand poses in ISL. For the initial research, gestures containing only one hand was considered. The solution described can be easily extended to two-handed gestures. In the next section of this paper, the related work pertaining to sign language translation is discussed.

II. RELATED WORK

There has been considerable work in the field of Sign Language recognition with novel approaches towards gesture recognition. A study of many different existing systems has been done to design a system that is efficient and robust than the rest.

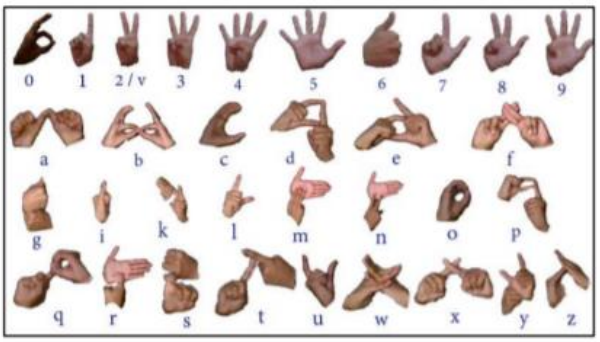

Fig. 1. Hand Poses in ISL

These hardware solutions provide good accuracy but are usually expensive and are not portable. Our system eliminates the need of external sensors by relying on an Android phone camera. Now for software-based solutions, there are coloured glove based and skin colour-based solutions. R. Y. Wang et al have used multi-coloured glove for accurate hand pose reconstruction but the sign demonstrator, while demonstrating the sign language, has to wear this each time. Skin colour-based solutions may use RGB colour space with some motion cues or HSV [9, 10, 11], YCrCb [12] colour space for luminosity invariance. G. Awad et al have used the initial

frames of the video sequence to train the SVM for skin colour variations for the further frames. But to speed up the skin segmentation, they have used Kalman filter for further prediction of position of skin coloured objects thus reducing the search space. Z. H. Al-Tairi et al have used YUV and RGB colour space for skin segmentation and the colour ranges that they have used handles good variation of people's races. After obtaining segmented hand image, A. B. Jmaa et al have used the rules defined in the hand anthropometry study of comparative measurements of human body for localizing and eliminating the palm.

They have then used the rest of the segmented image containing only fingers to create skin-pixel histogram with respect to palm centroid. This histogram is fed to decision tree classifier. In, from the segmented hand image, hand contour was obtained, which was then used for fitting a convex hull and convexity defects were found out. Using this, the fingers were identified and the angles between the adjacent ones were determined. This feature set of angles was fed to SVM for classification. [10]have used distance transform to identify hand centroid followed by elimination of palm and using angles between fingers for classification. Fourier Descriptors have been used to describe hand contours by has used RBF on these Fourier Descriptors for hand pose classification.

S. C. Agarwal et al have used a combination of geometric features (eccentricity, aspect ratio, orientation, solidity), Histogram of Oriented Gradients (HOG) and Scale Invariant Fourier Transform (SIFT) key points as feature vectors. The accuracy obtained using geometric features goes really low when number of hand poses increases. [8] has used Local Binary Patterns (LBP) as features. Our paper is mainly inspired from [9]. They have trained the k-NN model using the binary segmented hand images directly. This technique provides great speed when combined with fast indexing methods, thus making it suitable for real-time applications. But to handle the variations in hand poses, more data needs to be captured. With the use of grid-based features in our system, the model will become more user-invariant. For gesture recognition, hand centroid tracking is done which provides motion information [12].

Gesture recognition can be done using the Finite State Machine [10] which has to be defined for each gesture. C.Y. Kao et al [06] have used 2 hand gestures for training HMM that will be used for gesture recognition. They defined directive gestures such as up, left, right, down for their 2 hands and a time series of these pairs was input to the HMM for gesture recognition. C. W. Ng et al used a combination of HMM and RNN classifiers. The HMM Gesture recognition that we have used in our system is mainly inspired from [12]. They were using 5 hand poses and the same 4 directive gestures. This 9-element vector was used as input to the HMM classifier. Training of HMM was done using Baum-Welch re-estimation formulas.

III. IMPLEMENTATION

Using an Android smartphone, gestures and signs performed by the person using ISL are captured and their frames are transmitted to the server for processing. To make the frames ready for recognition of gestures and hand poses, they need to be pre-processed. The pre-processing first involves face removal, stabilisation and skin colour segmentation to remove background details and later morphology operations to reduce noise. The hand of the person is extracted and tracked in each frame. For recognition of hand poses, features are extracted from the hand and fed into a classifier. The recognised hand pose class is sent back to the Android device. For classification of hand gestures, the intermediate hand poses are recognised and using these recognised poses and their intermediate motion, a pattern is defined which is represented in tuples. This is encoded for HMM and fed to it.

The gesture whose HMM chain gives the highest score with forward-backward algorithm is determined to be the recognized gesture for this pattern. The server is set up using a Node server so as to handle the incoming REST requests from the Android Application. The Application captures the real time video into small frames and these frames are then sent to the deployed server. The Server also has a debug feature to see the different techniques being applied to the video and frames.

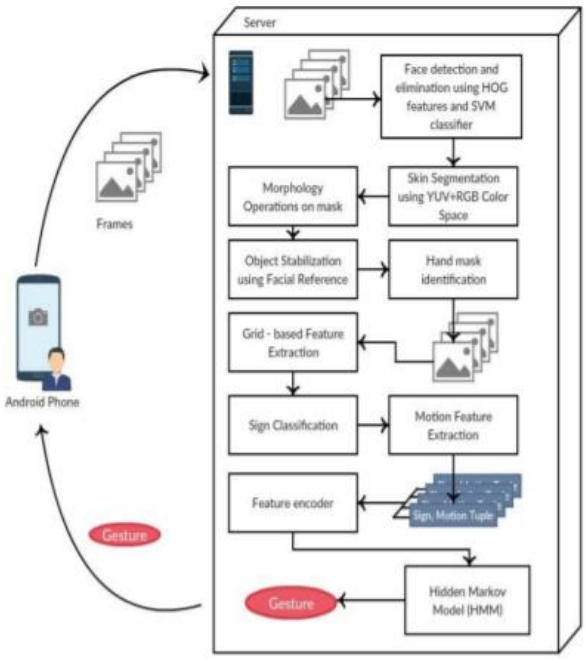

Fig. 2 Implementation

1. Datasets - For the digits 0 to 9 in ISL, an average of 1450 images per digit were captured. For letters in ISL excluding 'h', 'j' and 'v', about 300 images per letter were captured. For the 9 gesture-related intermediate hand poses such as ThumbsUp, SunUp, about 500 images per pose were captured. The dataset contains a total of 24,624 images. All the images consist of the sign demonstrator wearing a full sleeve shirt. Most of these images were captured from an ordinary webcam and a few of them were captured from a smartphone camera. The images are of varying resolutions. For training HMMs, 15 gesture videos were captured for each of the 12 one-handed pre-selected gestures defined in [1] (After, All The Best, Apple, Good Afternoon, Good Morning, Good Night, I Am Sorry, Leader, Please Give Me Your Pen, Strike, That is Good, Towards). These videos have slight variations in sequences of hand poses and hand motion so as to make the HMMs robust. These videos were captured from a smartphone camera and also involve the sign demonstrator wearing a full sleeve shirt.

2. Pre-Processing

2.1 Face Detection and Elimination- The hand poses and gestures in ISL can be represented by particular movement of hands, and facial features are not necessary. Also, face of the person creates an issue during hand extraction process. To resolve this issue, face detection was carried out using Histogram of Oriented Gradients (HOG) descriptors followed by a linear SVM classifier. It uses an image pyramid and sliding window to detect faces in an image, as described in [09]. HOG feature extraction combined with a linear classifier reduces false positive rates by more than an order of magnitude than the best Haar wavelet-based detector [10]. After detection of face, the face contour region is identified, and the entire face-neck region is blackened out.

2.2 Morphology operations- Morphology operations were performed to remove any noise generated after skin colour segmentation. There are 2 types of errors in skin colour segmentation:

2.2.1. Non-skin pixels classified as skin

2.2.2. Skin pixels classified as non-skin Morphology involves 2 basic sub-operations:

2.2.3 Erosion- Here, the active areas in the mask (which are white) are reduced in size 2.

2.2.4 Dilation- Here, the active areas in the mask (which are white) are increased in size Morphology Open operation handles the 1st error. It involves erosion followed by dilation. The 2nd error is handled by Morphology Close operation. This involves dilation followed by erosion 3).

3. Object Stabilisation using Facial reference.- For tracking hand motion accurately, a steady camera position is desired. Movement of the camera caused by shaky hands is common. If the sign demonstrator, that is, the person using ISL does not move his hand but the person capturing the video shakes his hand, false movements will get detected. This problem is tackled using object stabilisation. Under the assumption that a person's face is always included in the gesture video, the face of the sign demonstrator is tracked to stabilise hand movements. The tracker is initialised with the co-ordinates which were extracted from Face Detection before removing the face. The tracker detects the location of the facial blob and shifts the entire frame in the opposite direction of the detected motion of facial coordinates. The system uses the Kernelized Correlation Filter (KCF) tracker implemented in the OpenCV library to track the face in each frame. The operation of tracking is performed on the image before the face is blackened.

4. Hand extraction and tracking - As all ISL hand poses and gestures can be represented using hand movements, hand extraction and tracking is important part of the system. After pre-processing each frame, a black and white image is obtained, where white areas represent skin. These skin areas do not contain facial regions. They contain parts of the hand and other skin-like parts in the original image. Since each frame contains only 1 hand (the other hand is not visible) or both hands are touching each other, the only prominent contour present in the frame will be the person's hand. Thus, areas of all contours in the frame are calculated and the contour with the largest area is extracted.

5. Classification

5.1 Recognition of ISL Hand poses using k-NN- After plotting the graph it can be observed that data is organised into clusters. There is more than one cluster for the same hand sign. For classification, an algorithm was needed which can distinguish clustered data efficiently. K-Nearest Neighbours (k-NN) was found suitable for such distribution of data. Hand extracted from each frame of the live feed is subject to feature extraction using the previously discussed Gridbased fragmentation. This sample is then represented in an M*N dimensional feature map. Using Euclidean distance as a distance metric, k nearest samples, which are fitted previously in the classifier, are computed.

The distance computation can be performed using a brute force approach, wherein Euclidean distance between the sample and each fitted sample in the classifier is calculated and the k lowest distances are selected. Other optimal approaches include the KD-tree and Ball Tree. The most suitable approach for distance computation is dependent on size of the data. Brute force works well on small data size. For low dimensional data, KD Tree works well whereas for high dimensional data, Ball Tree works best. From these k nearest neighbours, the classifier selects the class occurring most frequently.

5.2 Temporal Segmentation- The gesture recognition module needs to be given video segments that correspond

to the gesture only. Without temporal segmentation, continuous gesture recognition is not possible. We have used a rule where in if the hand goes out of frame, it would mark the end of the current gesture and gesture recognition will be performed over the currently obtained frame sequence from gesture. When the hand again comes within frame, it would mark the beginning of the new gesture. Temporal segmentation is achieved using this rule.

IV. THE APPLICATION

The system is implemented as an Android application. The application uses the smartphone's camera to capture the sign language used by the person. The frames were captured at a rate of 5 frames per second. Each frame is continuously sent to a remote server. The processing is performed at the server-side. After each pose or gesture is classified, the result is sent back to the application which is displayed in the top portion. Currently, sockets are used to simulate a clientserver connection.

V. FUTURE WORK

The following model predicts 28 gestures which can be increased to include more and more gestures so that it can be more useful to the community.

VI. CONCLUSIONS

From the results, it can be concluded that the system presented in this paper is accurately able to track hand movements of the sign demonstrator using techniques such as (1)Object Stabilisation. (2)Face elimination. (3)Skin colour Extraction. (4)Hand extraction. It can classify all 28 hand poses in ISL with an accuracy of 96.4% . The system was also able to classify 10 gestures with an average accuracy of 98.23%. The approach uses an HMM chain for each gesture and a k-NN model to classify each hand pose.

The time required for recognition of hand pose is about 0.25s and that for gesture is 0.0048s. From the results, it can be concluded that the system can recognise hand poses and gestures in ISL with precision and in real-time. The system provides higher accuracy and faster recognition in sign language recognition than other approaches discussed in the literature. The approach discussed is inspired from various systems described in the Related Works section and utilises the pros discussed in their system to make the system more precise while classification. This approach is generic and can be extended to other singlehanded and two-handed gestures. The system presented in this paper can also be extended to other Sign Languages, if dataset satisfying the current requirements of the system is available.

REFERENCES

[1] Talking Hands.co.in, "Talking Hands," 2014. [Online]. Available: http://www.talkinghands.co.in/. [Accessed: 21- Jul- 2017].

[2] Mail Online, ''SignAloud gloves translate sign language gestures into spoken English," 2016. [Online]. Availablehttp://www.dailymail.co.uk/sciencetech/article3557362/SignAloudgloves- translate-sign-language-movements spoken-English.html. . [Accessed: 10- Feb- 2018].

[3] Alexia. Tsotsis, "MotionSavvy Is A Tablet App That Understands Sign Language," 2014. [Online]. Available: https://techcrunch.com/2014/06/06/motionsavvy-is-a-tablet-app that under stands- sign-language/. [Accessed: 10 – Feb- 2018].

[4] E. H. Miller, "A note on reflector arrays (Periodical stF. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al, "1.6. Nearest Neighbours – scikit-learn 0.19.1 documentation," 2011. [Online]. Available: http://scikitlearn. org/stable/modules/neighbors.html nearest-neighbor algorithms. [Accessed: 12- Sep- 2017]. of erbium-doped fiber amplifiers arrays (Periodical styleÑSubmitted for publication)," IEEE J. Quantum Electron., submitted for publication.

[5] R. Y. Wang and J. Popovic, "Real-Time Hand-Tracking with a Color Glove," ACM transactions on graphics (TOG), vol. 28, no. 3, 2009.

[6] C.W. Ng and S. Ranganath, "Real-time gesture recognition system and application," Image and Vision computing, 2002.

[7] A. Chaudhary, J. L. Raheja and S. Raheja, "A Vision based Geometrical Method to find Fingers Positions in Real Time Hand Gesture Recognition," JSW, pp. 861-869, 2012.

[8] L. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Process," Inequalities III: Proceedings of the Third Symposium on Inequalities, ssvol. 3, pp. 1-8, 1972.

[9] M. Hrúz, J. Trojanová and M. Železný, "Local Binary Pattern based features for Sign Language Recognition," Pattern Recognition and Image Analysis, 2012.

[10]. L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, no. 2, February 1989.

[11] L. Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, November 2008.

[12] M. A. Rahaman , M. Jasim, M. H. Ali and M. Hasanuzzaman, "RealTime Computer Vision-Based Bengali Sign Language Recognition," 17th

International Conference on Computer and Information Technology (ICCIT), 2014..