RESEARCH ARTICLE OPEN ACCESS

Continuous Sign Language Recognition Using LSTM and Media Pipe Holistic

Madhura Mirikar*, Komal Singh**, Prof. Dr. Sampada Dhole***

*(Department of Electronics and Telecommunication Engineering,
Bharati Vidyapeeth's College of Engineering for Women, Pune
Email: madhura.mirikar@gmail.com)

**(Department of Electronics and Telecommunication Engineering,
Bharati Vidyapeeth's College of Engineering for Women, Pune
Email: komalsingh7102000@gmail.com)

***(Department of Electronics and Telecommunication Engineering,
Bharati Vidyapeeth's College of Engineering for Women, Pune
Email: sampada.dhole@bharatividyapeeth.edu)

--************************----------------------------------

Abstract: Speech impairment is a communication disorder that impacts an individual's ability to speak fluently, correctly, or with clear resonance or tone. This paper presents a robust system to aid in communicating with those suffering from speech impairment. Advancements in technology have led to the development of innovative approaches for gesture recognition. The main purpose of this paper is to introduce an improved method for Sign Language Recognition (SLR) using Mediapipe and Long Short-Term Memory (LSTM), an artificial neural network. The proposed system performs real-time gesture recognition on 5 signs from the American Sign Language (ASL) and gives precise, accurate, and efficient results with an average accuracy of up to 99%

Keywords **—Sign Language Recognition, Mediapipe, Long Short-Term Memory, American Sign Language**

--************************----------------------------------

I. INTRODUCTION

Speech disorders affect roughly 11.5% of the US population and approximately 18.5 million individuals worldwide have a speech, voice, or language disorder. Figures of WHO state that nearly 466 million people that comprise approximately 5% of the World's population are with such disabilities and out of which 35 million are children [6]. Over 5% of the world's population – or 430 million people – require rehabilitation to address their 'disabling' hearing loss (432 million adults and 34 million children). It is projected that by 2050 nearly 2.5 billion people are projected to have some degree of hearing loss and at least 700 million will require hearing rehabilitation. [7] as shown in Fig 1.

Sign language is a means of communicating with those who have hearing and vocal disorders. It is a non-verbal visual language that makes use of both 'manual features' (hand shape, position, orientation, and movement) and linguistically termed 'non-manual' features consisting of the face (eye gaze, mouthing/mouth gestures, and facial expression) and the upper body posture (head nods/shakes and shoulder orientation). [1] [2] Sign language varies from different countries to regions i.e. there is no universal sign language. In our project, we worked on American Sign Language which is used predominantly by the dumb and deaf communities in America and some parts of Canada.

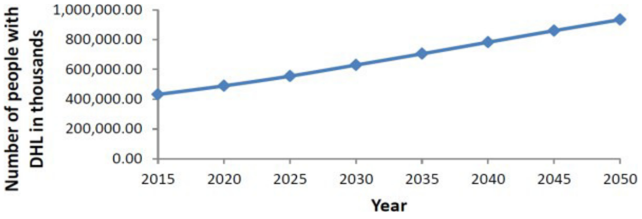


Fig 1. WHO estimates on Disabling Hearing Loss (DHL)

Formerly, Sign Language Recognition systems were classified in two ways, sensor based system [3] [4] or computer-vision based system. Sensor based techniques proved to be restrictive and uncomfortable for signers, moreover, the equipment turned out to be comparatively expensive. Computer-vision based systems are budget-friendly and mobile as compared to sensor based techniques. These systems use a camera and image-processing techniques to validate actions/gestures. [5] But, all these solutions are restricted to run in platforms equipped with high-level processors. To overcome these drawbacks, our proposed methodology uses MediaPipe, that includes optimised face, hands, and pose components for holistic tracking, allowing the model to identify hand and body poses as well as face landmarks at the same time [8] and Long Short-Term Memory (LSTM), an artificial neural network which is able to look at long sequences of inputs without increasing the network size to develop a sign language recognition system that is accurate, cost-effective, portable, and easy to deploy.

II. RELATED WORK

Existing SLR systems are based on 2D video camera [17], colored gloves [2], sensor-gloves [25], [22], etc. Recently, the SLR research is shifting to a novel 3D environment using depth cameras/sensors [19], [29]. Most of the work on the recognition of sign gestures are based on HMMs, Artificial Neural Networks (ANN) and rule-based modelling techniques.

The authors in [19] have used a view-based approach in developing a continuous-SLR system using a single video camera. In their first stage, the camera, mounted on a disk observed the signer whereas in the second stage the signer was observed by a cap mounted camera. Vision-based skin-colour modelling was used for hand segmentation and hand blob extraction. Next, the authors have extracted a 16- dimensional feature vector by doing hand blobs analysis that includes positions, angular, area and length features. Training was carried out on 384 and 400 ASL sentences for desk-mounted and capmounted based systems, respectively. The system was tested on 94 and 100 ASL sentence using HMM with a vocabulary of 40 signs where the accuracy of 74.5% and 97.8% were recorded with desk-based and cap-based systems, respectively. In [3], the authors have proposed a continuous-SLR system for German Sign Language (GSL) using a video camera. The authors have used coloured gloves for data acquisition and hand segmentation. The recognition process was carried out using HMM-based language modelling with uni-gram and bigram models on two different vocabularies of 52 and 97 signs. They have extracted hand positions, angular and distance based features, which were fed directly to HMM, where the accuracy of 95.4% and 93.2% were recorded on 52 and 97 lexicon signs with bi-gram language models, respectively. However, the system had restrictions on signer's clothing and required a uniform coloured background for correct segmentation of hands. A framework for the continuous-SLR system using three orthogonal cameras was proposed by Vogler et al. [16] to capture 3D hand movements. The authors have extracted 8-dimensional feature vector that consists of 3D hand positions, velocity and eigenvalues of the positions covariance matrices. The recognition process was performed using parallel-HMM on 99 ASL sentences with an accuracy of 84.84% that outperforms the conventional HMM-based recognition.

Li et al. [15] proposed a model-based framework for segmentation and recognition of continuous SLR using the video sequence. The authors presented three different approaches for endpoint localization of gestures that include multi-scale search, Dynamic Time Warping (DTW) and dynamic programming. They extracted the hand contour as a feature vector. The system was tested on 12 continuous gestures, where a recognition rate of 82% was recorded using early-decision dynamic programming with correlation and mutual information based similarity measures. A framework for segmentation, tracking and modelling of hand shapes from video sequences was proposed in [18] using probabilistic skin colour analysis, forward-backwards prediction and affine-invariant modelling, respectively. It offers a compact and descriptive representation of the hand configuration. The hand-shape features have been extracted using the affine modelled hand that was used to construct an unsupervised set of sub-units which constitute the signs.

Gao et al. [6] have proposed an SLR system for Chinese Sign Language (CSL) using data gloves and three position trackers to extract the hand appearance and position. The authors have extracted 48-dimensional feature vector that includes hand shape, position and orientation vector. A modified k-means clustering algorithm was used with DTW based distance measuring technique to cluster the transition movements between two signs. The system was tested on 750 CSL sentence with a vocabulary size of 5113 signs, where the accuracy of 90.8% was recorded. It was assumed that the transition movement between two signs is always similar in different sentences. However, such transitions vary in real-world applications. The authors in [11] have used similar data gloves based approach for Arabic Sign Language recognition system. They have used a modified version of k-NN algorithm for classification of 40 signed sentences. However, the evaluation was performed in user-dependent mode. Another digital glove based study can be found in [16], where the authors proposed a scalable HMM system for continuous-SLR by training a single universal transition model. Yang et al. [12] have proposed a continuous SLR framework using the Kinect. The authors have also proposed a low complexity level building algorithm for computing the likelihood of HMM. Six 3D skeleton features were extracted from Kinect SDK. The system has been tested on 100 CSL sentences on a vocabulary of 21 signs with an error rate of 12.20% when tested with the HMM-based classifier. Recently, the authors in [10] have proposed calibration of Leap Motion and Kinect sensors for the recognition of gestures in ISL. They have recorded 50 isolated sign gestures of ISL and tracked the 3D positions of the finger and hand movements. Angular features were extracted for the recognition purpose that has been performed using HMM classifier. Similarly, [12], [13] used 3D text segmentation and recognition methodology using Leap motion sensor. The authors recorded 3D sentences in the air over the Leap motion view field.

III. PROPOSED METHODOLOGY

In this section, we present our LSTM based neural network architecture for gesture recognition used for Sign Language Recognition using Mediapipe Holistic. The flow diagram of the framework is depicted in Fig. 2.

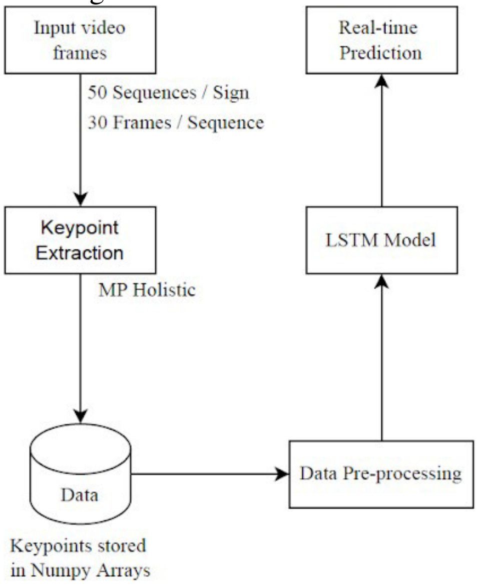


Fig 2. System framework for Sign Language Recognition

A. Data Collection and Key-Point Extraction

Using the camera of the device and Computer Vision library, 50 videos of each 1-2 seconds are taken for each of the 5 Signs. Each video has a frame length of 30 i.e. there are 30 frames per video and 50 videos per sign. Each frame is converted from BGR to RGB since OpenCV interprets BGR format by default. Using Mediapipe Holistic, keypoints are extracted from the frames. There are a total of 468 face keypoints, 33 pose keypoints and 21 keypoints for each hand respectively.

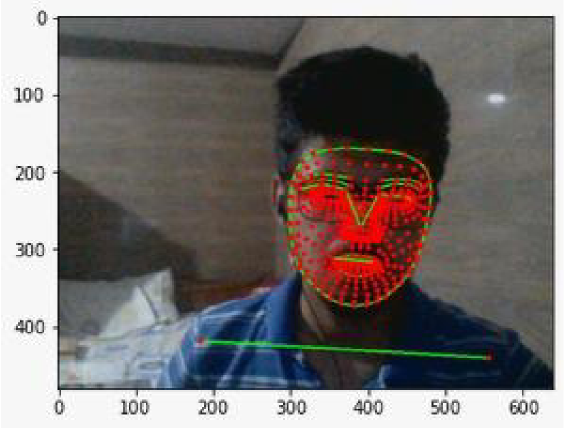


Fig 3. Face Landmarks using MP Holistic

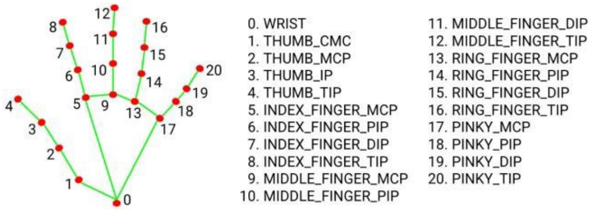


Fig 4. Hand Landmarks using MP Holistic

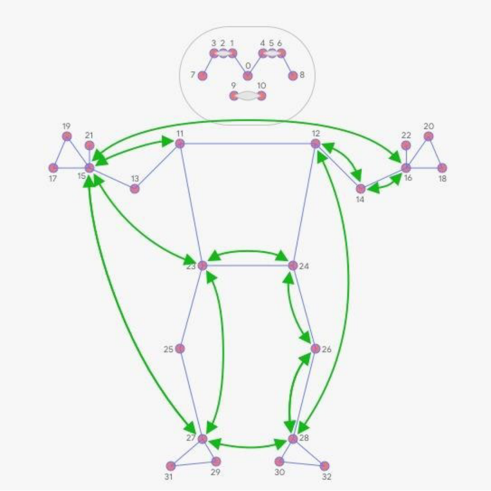


Fig 5. Pose Landmarks using MP Holistic

B. Pre-Processing Data

Each keypoint for Face and Left_Hand as well as Right_Hand has 3 landmarks whereas, for Pose keypoints, there are 4 landmarks. All these landmarks are extracted and stored in Numpy Arrays for future processing. The total number of landmarks would be (468 * 3) + (33 * 4) + (21 * 3) + (21 * 3) = 1662. The arrays of landmarks are stored in different local folders per sign. A label map was created for each of the five signs.

label_map = {'hello': 0, 'thanks': 1, 'iloveyou': 2, 'please': 3, 'house': 4}

The data was split into training and testing parts using Scikit-learn's train_test_split utility where 20% of the data was reserved for validation.

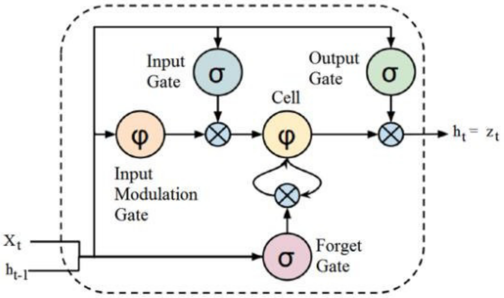
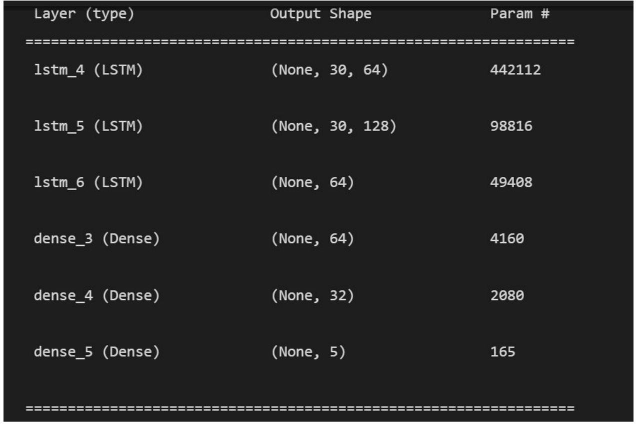


Fig 6. LSTM Architecture

LSTM is a special kind of recurrent neural network that is capable of learning long term dependencies in data. This is achieved because the recurring module of the model has a combination of four layers interacting with each other. An LSTM module has a cell state and three gates which provides them with the power to selectively learn, unlearn or retain information from each of the units. The cell state in LSTM helps the information to flow through the units without being altered by allowing only a few linear interactions. Each unit has an input, output and a forget gate which can add or remove the information to the cell state. The forget gate decides which information from the previous cell state should be forgotten for which it uses a sigmoid function. The input gate controls the information flow to the current cell state using a point-wise multiplication operation of ‘sigmoid’ and ‘tanh’ respectively. Finally, the output gate decides which information should be passed on to the next hidden state.

Our proposed structure for LSTM model used for Sign Language Recognition is depicted in Fig 7.



```
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_4 (LSTM)               (None, 30, 64)            442112

 lstm_5 (LSTM)               (None, 30, 128)           98816

 lstm_6 (LSTM)               (None, 64)                49408

 dense_3 (Dense)             (None, 64)                4160

 dense_4 (Dense)             (None, 32)                2080

 dense_5 (Dense)             (None, 5)                 165

=================================================================
```

Fig 7. LSTM Model proposed for SLR

This model is trained with Activation functions ReLu (Rectified Linear Unit) and SoftMax for the last Dense layer. The optimizer used in our model is Adam optimizer.

The data was trained for 100 epochs and achieved up to 98.68% accuracy.

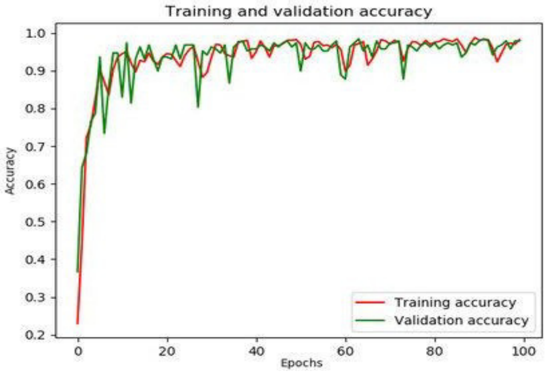


Fig 8. Training and Validation Accuracy

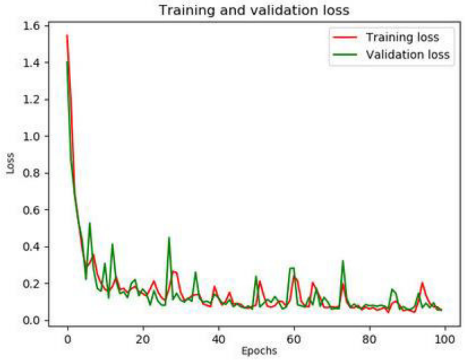
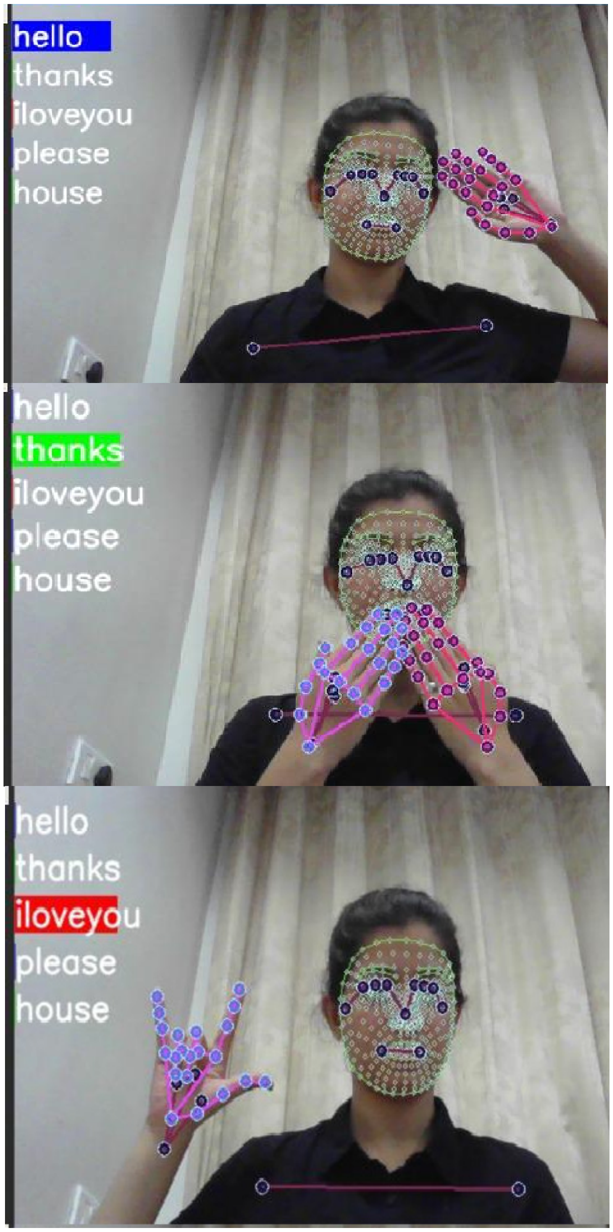


Fig 9. Training and Validation Loss

IV. RESULTS

Thesignsarefurtherpredictedinreal-timeusingcomputervisionandthe outputsare asfollows:



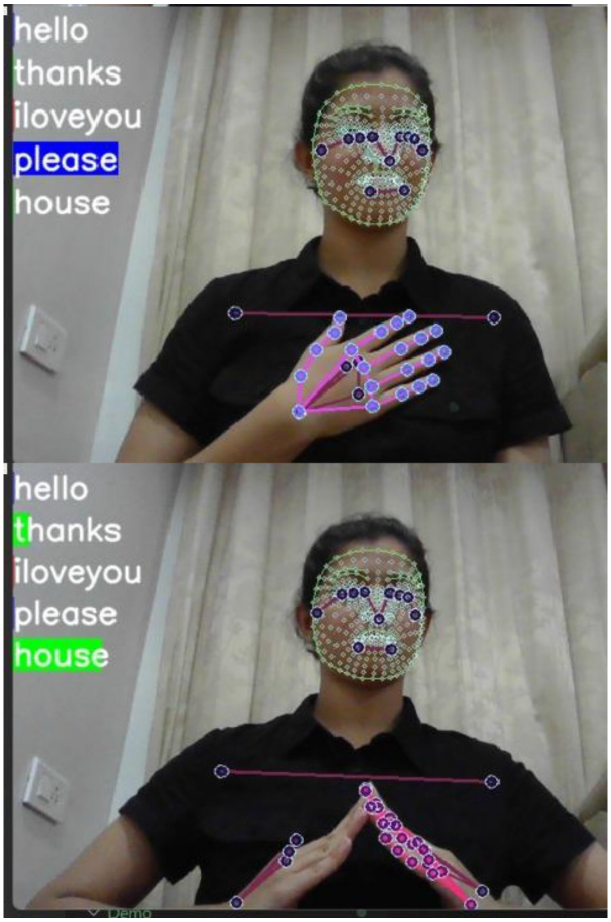


Fig 10. Implemented Outputs

Theclassificationhasachievedanaverageclassificationaccuracy of 98.68%. It means all of the gesture videos in thetest set have been classified into right classes. If there is anymisclassification present may be reason behind that is, theinconsistency of spatial andtemporal features due to thesimilarityinthe appearanceof gestures.

V. CONCLUSION

In this paper, we have presented a novel framework forcontinuous-SLRusingMediapipeHolisticandLSTMNetwork. We trained a dataset of 5 signs each containing 50videos and each video was comprised of 30 frames. 1662Keypointswereextracted fromeachframeandtheprocesseddata was stored in Numpy arrays. The dataset was trainedusingLSTMNeuralNetworkand20%of theDatawasreserved for testing. Further, gestures were predicted in real-timeandourmodelachievedtheaccuracyof98.68% .

VI. FUTURESCOPE

Whilethecommunitycontinuestodiscusstheneedfor including non-manual features, few have actually done so.Thosewhichhave,concentratesolelyonthefacialexpressions of sign. There is still much to be explored in theveinsofbodypostureorplacementandclassifier(handshape)combinations.Finally,tocompoundallthesechallenges,thereisthe issueofsignerindependence.

While larger data sets are starting to appear, few allow truetestsofsignerindependenceoverlongcontinuous sequences. Maybe this is one of the most urgent problems inSLR that of creating data sets which are not only realistic,butalsowellannotated tofacilitatemachinelearning.

This model can be deployed in mobile applications, web-applications and even in chat-bots. Number of signs can beincreased inthefuture.

ACKNOWLEDGMENT

We would like to thank the open-source community on theInternet,inparticularthetutorsatTensorFlowandMediaPipe, whose help was indispensable to our learningprocess.

REFERENCES

[1] Moeslund, Thomas B.; Hilton, Adrian; Krüger, Volker; Sigal, Leonid(2011). “Visual Analysis of Humans ‖ Sign Language Recognition.” ,10.1007/978-0-85729-997-0(Chapter 27), 539–562. doi:10.1007/978-0-85729-997-0_27

[2] Koller, Oscar; Forster, Jens; Ney, Hermann (2015). “Continuous signlanguagerecognition:Towardslargevocabularystatisticalrecognition systems handling multiple signers. Computer Vision andImage Understanding”,141(), 108–125.doi:10.1016/j.cviu.2015.09.013

[3] K. Assaleh, T. Shanableh, M. Fanaswala, F. Amin, H. Bajaj, et al.“Continuousarabicsignlanguagerecognitioninuserdependentmode.JournalofIntelligentlearningsystemsandapplications”,2(01):19,2010.

[4] B. Bauer, H. Hienz, and K.-F. Kraiss. “Video-based continuous signlanguage recognition using statistical methods.”, 15th InternationalConference on Pattern Recognition, volume 2, pages 463–466. IEEE,2000.

[5] B. Bauer and K. Karl-Friedrich. “Towards an automatic sign languagerecognition system using subunits.”, International Gesture Workshop,pages64–75.Springer,2001.

[6] S. K. Behera, D. P. Dogra, and P. P. Roy. “Analysis of 3d signaturesrecordedusingleapmotionsensor.MultimediaToolsandApplications”,pages1–26,2017.

[7] H. Cooper, B. Holt, and R. Bowden. “Sign language recognition. InVisualAnalysisofHumans”,pages539–562.Springer, 2011.

[8] W. Gao, G. Fang, D. Zhao, and Y. Chen.“Transitionmovementmodels for large vocabulary continuous sign language recognition.”,InInternationalConferenceonAutomatic FaceandGestureRecognition,pages 553–558. IEEE,2004.

[9] R. Haldar and D. Mukhopadhyay. “Levenshtein distance technique indictionary lookup methods: An improved approach.” arXiv preprintarXiv:1101.1232,2011.

[10] W.KongandS.Ranganath.“Towardssubjectindependentcontinuoussignlanguagerecognition:Asegmentand mergeapproach.PatternRecognition”,47(3):1294–1308,2014.

[11] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra. “Coupled hmmbased multi-sensor data fusion for sign language recognition”. PatternRecognition Letters,86:1–8,2017.

[12] P.Kumar,R.Saini,P.Roy,andD.Dogra.“Studyoftextsegmentationandrecognitionusingleapmotionsensor”,IEEESensorsJournal,2016.

[13] H.LiandM.Greenspan.“Model-basedsegmentationandrecognitionofdynamicgesturesincontinuousvideostreams.”,PatternRecognition, 44(8):1614–1628,2011.

[14] K.Li,Z.Zhou,andC.-H.Lee.“Signtransitionmodelingandascalable solution to continuous sign language recognition for real-

world applications.”, ACM Transactions on Accessible Computing,8(2):7,2016.

[15] Z.Zafrulla,H. Brashear,T.Starner,H.Hamilton, andP.Presti.“Americansignlanguagerecognitionwitht hekinect.”,In13thinternational conference on multimodal interfaces, pages279–286.ACM, 2011.

[16] HandTrackingandRecognitionWithOpenCV[Online]. Availablehttp://simena86.github.io/blogl2013/08/12I hand-trackingandrecognition-with-opencvl

[17] Mokhtar M. Hasan & Pramod K. Misra,“HSV Brightness FactorMatching for Gesture Recognition System.”, International Journal ofImageProcessing(IJIP),Volume(4):Issue(5)

[18] Xingyan. Li. "Vision Based GestureRecognition System with HighAccuracy".DepartmentofComputerScience,The UniversityofTennessee, Knoxville,TN37996-3450,2005

[19] Zieren,J.,Kraiss,K.F.,“Robustperson-independentvisualsignlanguage recognition.” In: Procs. of IbPRIA, Estoril, Portugal, pp.520–528 (7–9June2005)