This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2909837, IEEE Sensors Journal

A Modified-LSTM Model for Continuous Sign Language Recognition using Leap motion

Anshul Mittal, Pradeep Kumar, Partha Pratim Roy, Raman Balasubramanian and Bidyut B. Chaudhuri

***Abstract*—Sign language facilitates communication between hearing impaired peoples and the rest of the society. A number of Sign Language Recognition (SLR) systems have been developed by researchers but they are limited to isolated sign gestures only. In this paper, we propose a modified LSTM model for continuous sequences of gestures or continuous SLR that recognizes a sequence of connected gestures. It is based on splitting of continuous signs into sub-units and modeling them with neural networks. Thus, the consideration of different combination of sub-units is not required during training. The proposed system has been tested with 942 signed sentences of Indian Sign Language (ISL). These sign sentences are recognized using 35 different sign words. The average accuracy of 72.3% and 89.5% have been recorded on signed sentences and isolated sign words, respectively.**

Keywords: Sign Language Recognition; Depth Sensors; Deep Neural Networks; Leap Motion Sensor.

I. Introduction

Sign language is one of the communication media for the hearing-impaired people. It is a non-verbal visual language that is characterized by manual and non-manual signs [5], [8], [16]. Non-manual signs consist of facial expressions, mouth and head movements etc., whereas manual signs consist of hand and finger movements, hand orientation and gesture, etc. Communication through sign language varies in different countries and thus they lack the uniformity, e.g. Indian Sign Language (ISL) in India [10], [11], American Sign Language (ASL) in America [27], etc. The structure of the sign language varies in spatial and temporal information, which is drastically different from the spoken language where one word follows another. However, the general configuration of the sign language sentence consists of: time, location, person and predicate, as discussed in [2].

Sign Language Recognition (SLR) aims to develop an assistive system that automatically converts an input sign into corresponding text/speech. SLR systems are helpful in bridging the communication gap between the hearing impaired and the rest of the society. Thus, such systems open a new way for Human-Computer-Interaction (HCI) based applications. Researchers have developed a number of SLR systems that work well for isolated words but fail to recognize and translate the continuous sequence of gestures. The primary challenge in developing a continuous SLR system is to find a modelling paradigm that can capture the sign gestures and corresponding language. In speech recognition, the problem is taken care of considering language modelling by phoneme units that appear sequentially. However, the same principle is applicable in continuous SLR systems where instead of words or phonemes, sequences of signs are available. Vogler et al. [23] developed a continuous SLR system for American Sign Language (ASL) sentences using three orthogonally positioned cameras to alleviate the problems caused by occlusion and unconstrained movements. The recognition process used Hidden Markov Model (HMM) with a vocabulary of 53-signs. The system was tested on 97 sign sentences with and without bigram language modelling, which yielded a recognition rate of 92.11% and 95.83%, respectively. Similarly, Starner et al. [21] developed a video-based real-time continuous SLR system for ASL sentences using a single camera with a vocabulary of 40 signs. Here the recognition process was carried out using HMM classifier. However, the SLR systems that are developed using single video camera suffers from problems due to (a) occlusion of signing fingers and hand movements, (b) the signer may not always remain in front of the camera, and (c) loss of the depth because of 2D nature of the single camera [3].

With the help of depth enabled sensors such as Leap motion [19] and Microsoft Kinect [29], which provide 3D point cloud of the observed scene improves the understanding of input data. Kinect is used to interpret the movement of the whole body whereas Leap motion provides real-time tracking of finger and hand movements. These devices help in overcoming the problems occurred in SLR systems captured by 2D video cameras as discussed earlier. In present work, we have used the Leap motion sensor to acquire the 3D data of hand and finger movements for developing a continuous SLR by individual modeling of each sign gesture. Most of the existing continuous SLR systems use HMMs which offer implicit segmentation of a sign gesture sequence into its constituent signs. However, a Markov chain is conditioned on a fixed window. With the advanced machine learning technologies like Recurrent Neural Network (RNN), it makes the generation and recognition of character from the history of the characters. In this paper, we propose a modified Long Short Term Memory (LSTM) architecture by introducing a Reset (R) gate which helps in segmenting the continuous sign sequence as well as in improving the recognition performance. The Convolutional Neural Network (CNN) has been used to extract the spatial features from the signed sequences which are then modelled by the modified LSTM model for recognition. Thus, the main contributions of the paper are as follows:

- Firstly, we present a novel framework for tracking and recognizing continuous Indian Sign Language (ISL) sen-

Anshul Mittal, Pradeep Kumar, Partha Pratim Roy and Raman Balasubramanian are with the Department of Computer Science Engineering, Indian Institute of Technology, Roorkee, India.

Bidyut B. Chaudhuri is with Techno India University, Kolkata, India.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2909837, IEEE Sensors Journal

tences based on modelling of sign sub-units.
- Secondly, we present a modified Long Short-Term Memory (LSTM) classifier for the recognition of continuous signed sentences using sign sub-units.
- Finally, the performance of the system is compared with traditional LSTM and state-of-the-art architectures.

Rest of the paper is organized as follows: Section II presents an overview of the existing work in the field of continuous-SLR. Section III describes our proposed methodology for implementing continuous SLR system which includes pre-processing, feature extraction, training and testing steps in details. In Section IV, we have discussed the dataset and compiled the results from different experiments. Finally, concluding remarks along with the possibilities of future work are given in Section V .

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 [21] 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 cap-mounted 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 bi-gram 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. [24] 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 [20] 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 [22] 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. [26] 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 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 continuous SLR using Leap motion sensor.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2909837, IEEE Sensors Journal

The flow diagram of the framework is depicted in Fig. 1, where the Leap motion sensor is used to acquire the sign inputs. The

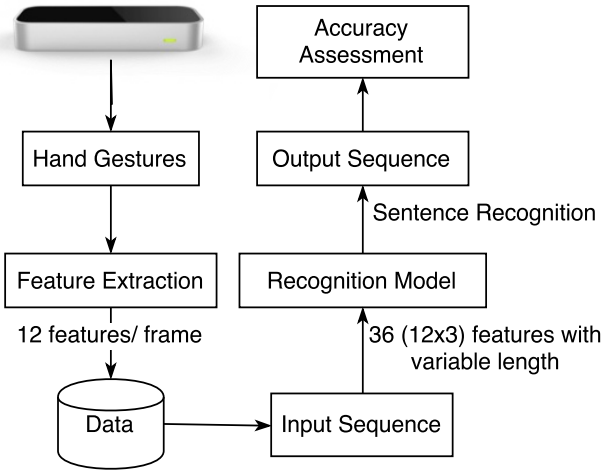

Fig. 1. Proposed framework for continuous SLR using Leap motion sensor.

sensor has two monochromatic Infrared (IR) cameras and three infrared Light Emitting Diodes (LEDs) that generate IR light. The cameras of the sensor are capable of generating almost 120-200 frames per second (fps) of reflected data [18]. We set this value to 120 fps. The sensor can observe a roughly hemispherical area of 1 cubic meter. A pictorial representation of the Leap motion Cartesian coordinate system is shown in Fig. 2, where x- and z-axes lie in the horizontal plane. The y-axis is vertical, with positive values increasing upwards [4]. We have extracted the 3D positions of the fingertips using the sensor's Application Programming Interface (API). The extracted features were then stored and processed for prediction.

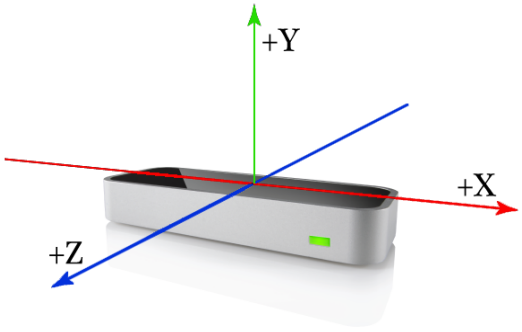

Fig. 2. Pictorial representation of the Leap motion Cartesian coordinate system.

A. Pre-processing and Feature Extraction

Data acquired by Leap Motion device consists of redundant information such as, joints of fingers and wrist. These redundant information are removed and only relevant data are retained. To manage the variation in hand spans, palm size and movements, a few preprocessing steps are performed. In the following subsection, we provide the details of the preprocessing and feature extraction steps that have been used before modeling of continuous sign gestures.

1) Feature Extraction: For each signer producing the sign gesture, we have extracted 12 dynamic features, i.e., $(f_1 ... f_5 ..... f_{12})$ for both hands. Each $f_i = (x, y, z)$ where x,y,z are co-ordinates of fingertip in 3D space. These features are extracted independently from data acquired by leap motion device. These features depicted in Fig. 3 are marked with red color. Features f_6 and f_{12} are palm centers for left and right hand, respectively.

In this work, a gesture of sign at a time t is represented by S_t where $S_t = [f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}]$ where each f_i is a 3D data point at time t as shown in Fig. 3.

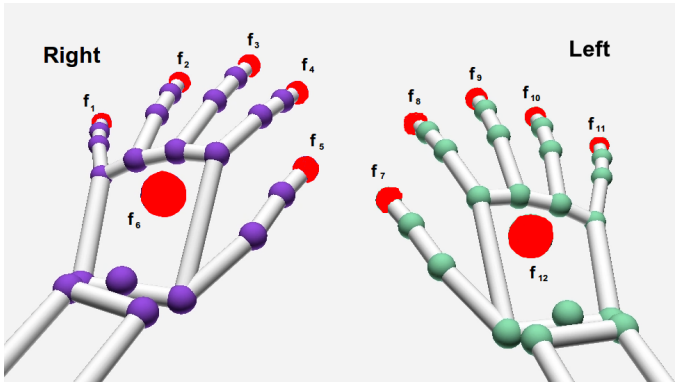

Fig. 3. Features extracted using Leap motion (red color).

2) Size Normalization: In this step, the features f_i in sign sequences S are scaled between $[-1, 1]$ to make them uniform. This scaling is performed by normalizing each dimension by its maximum value within the sequence. It helps in removing the signer's specific conditions like hand spans, palm size and movements from the sign language sequences.

B. Convolutional Neural Network (CNN)

In machine learning, a CNN is a class of deep, feed-forward artificial neural networks. CNN uses a variation of multilayer perceptrons designed to attain minimal preprocessing [14]. These networks are also known as shift or space invariant due to shared-weights architecture [28]. Typically, a CNN consists of an input and an output layer in combination of multiple hidden layers which are convolutional, pooling, fully connected and normalization layers. In this work, we have used 2-dimensional CNN (2D CNN) for feature extraction. A pictorial illustration of this step is shown in Fig. 4, where twelve 3D features (S_t) at time t are processed with three kernel of size $3*5$ that returns three feature maps of size $1*8$. Next, these feature maps are fed into the proposed modified LSTM network for recognition.

C. Long Short Term Memory (LSTM) Network

A conventional LSTM consists of input gate, output gate, forget gate and memory blocks in the recurrent layer. The memory blocks contain memory cells which store the temporal

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2909837, IEEE Sensors Journal

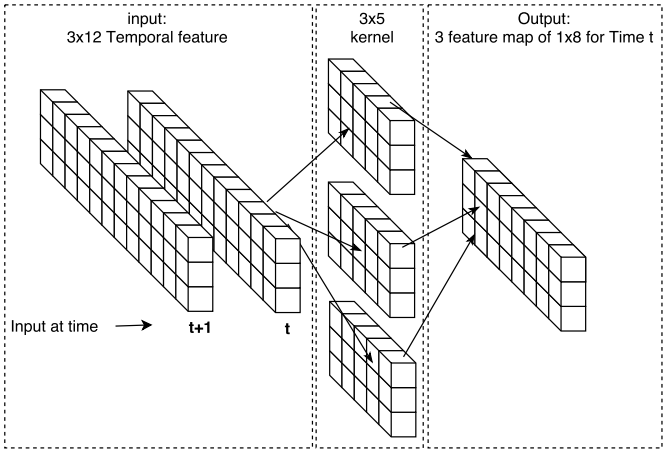

Fig. 4. 2D CNN based feature extraction at time t.

state of the network using self-connections, in addition to the gates to manipulate the flow of information. The input gate controls the input of flow of activations into the memory cells. The output gate controls the flow of output of cell activations to the rest of the network. The forget gate, through the self-recurrent connection to the cell, scales the internal state of the cell before adding it as input to the cell, hence adaptively forgetting the cell's memory content. An LSTM network computes a mapping from an input sequence $A = (A_1, ..., A_T)$ to an output sequence $B = (B_1, ..., B_T)$ by calculating the network unit activations iteratively from t = 1 to T.

D. Modified-LSTM Network

We present a modified four gated LSTM cell with 2D CNN for sign sentence recognition. The architecture is shown in Fig. 5. Three basic gates operating at time t are the input gate

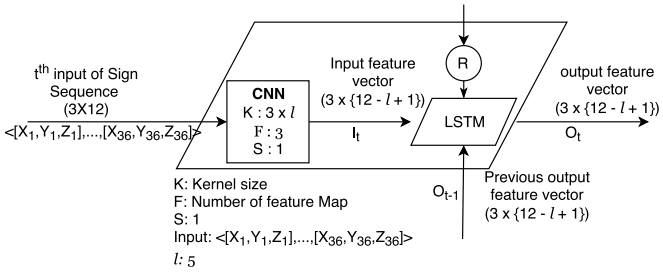

Fig. 5. Proposed modified LSTM architecture with a $RESET(R)$ gate.

which takes input I_t, output gate which gives the output O_t and forget gate which takes the output at time $t-1$. A 4^{th} gate is the $RESET(R)$ gate which resets the memory cell of the LSTM for every **True (i.e. 1)** as input in R (i.e. State of LSTM cell at time $t = 0$). The LSTM predicts a class number or a word with the label ($) which refers 'no-class'. Here, we introduce this extra-label ($) to capture the transition between two consecutive signs. Therefore, when ($) is encountered, the network will send a signal **True** to RESET gate to reset the memory of the LSTM, which removes any influence of the previously predicted word. This resetting helps in segmenting the individual sign words from the input sequence and also helps in improving the recognition performance. For better understanding, a pictorial representation of both basic and modified LSTM units is depicted in Fig. 6.

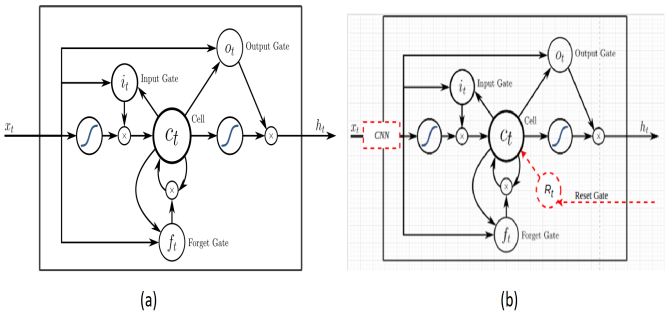

Fig. 6. Pictorial representation of LSTM units: (a) basic LSTM (b) modified LSTM with Reset functionality.

The LSTM network with N memory cells form an input sequence $S = (S_1, ..., S_T)$ and an output sequence $B = (B_1, ..., B_T)$. Each S_i from S consists of L features, each of 3 dimensions i.e. x,y and z. Thus, the kernel Q is of size $3 * L$ which calculates the network unit activation using the following equations iteratively from $t = 1$ to T:

$K_{t,j}$ is the j^{th} input feature at time t for the input gate. The process is defined in (1)-(7), where the W, b aand σ are the weight matrices, bias vectors, and logistic sigmoid function, respectively. The terms i, f, o and c are the input, forget, output and cell activation vectors, respectively. In this paper, $tanh$, and ψ in the network are used as output activation functions.

$$K_{t,j} = Q * (S[t,j], S[t,j+1]...S[t,j+l-1]) \quad (1)$$

$$i_t = \sigma(W_i K_t + W_{im} m_{t1} + W_{ic} c_{t1} + b_i) \quad (2)$$

$$f_t = \sigma(W_f K_t + W_{fm} m_{t1} + W_{fc} c_{t1} + b_f) \quad (3)$$

$$c_t = f_t c_{t1} + i_t g(W_c K_t + W_{cm} m_{t1} + b_c) \quad (4)$$

$$o_t = \sigma(W_o K_t + W_{om} m_{t1} + W_{oc} c_t + b_o) \quad (5)$$

$$m_t = o_t h(c_t) \quad (6)$$

$$B_t = \psi(W_{Bm} m_t + b_B) \quad (7)$$

1) Traning and Testing: Our model consists of three layers of modified LSTM with Rectified linear unit (ReLU) activation followed by dense layer and one Softmax layer for classification into multiple classes. We used Adadelta optimizer which does not requires manual tuning of learning rate and has been found to be robust against noisy gradient. The model is first trained using isolated sign gestures and then fine-tuned with continuous gestures for efficient modeling of the framework. The continuous signs are generated from the isolated gestures by adding a variable length transitions states/gesture.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2909837, IEEE Sensors Journal

Fig. 7 depicts the training model which consists of three LSTM layers and a fully connected layer with Softmax activation. The term 'C'' in the decision box denotes the class number for the input sequence. It can be seen from Fig. 7 that the output is recorded at each input frame for a given sequence. If C is not associated with the class number of transition state ($), the LSTM memory unit is preserved until 'C'' gives a class number of transition state ($) with probability > 0.5 at input frame A_t at time t. At that time a positive feed is given at Gate R for each LSTM layer to reset the value of memory cell to remove the effect of input frames till A_t. After repeating the process for the entire sequence, Levenshtein distance [7] is used to calculate the distance between predicted and actual vector of words by considering the vector of words as a string. The store output unit represents the predicted word string $(W_1, W_2, .., W_i)$, where W_i is the i^{th} predicted word.

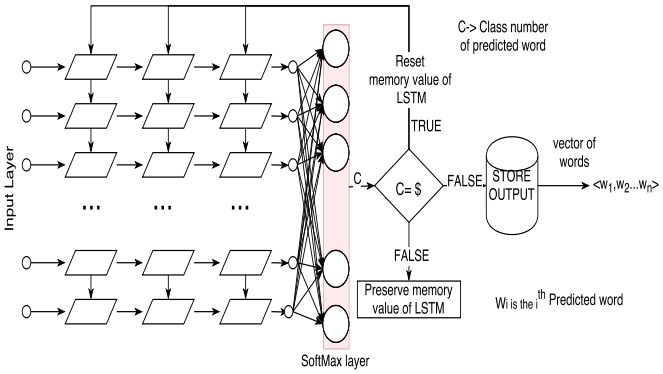

Fig. 7. Modified LSTM model proposed for continuous Sign Language Recognition.

IV. Dataset Description and Results

Here, we present the details of the dataset that has been recorded using the proposed SLR framework. Next, the recognition of continuous sign language has been presented. Finally, recognition results of isolated sign words have been shown.

We have enrolled six participants for sign language data collection. All participants are the students of a hearing impaired school situated at IIT Roorkee, India. The dataset consists of 35 isolated sign words. Each sign word has been repeated at least 15 times by every signer. Therefore, a total of 3150 (35*15*6) sign words were recorded. A description of all sign words is presented in Table I, where '$' sign represents the transition gesture (i.e. when one user is changing from one word to other) between two continuous signs while performing sign sentences.

TABLE I
Sign words available in the dataset

friend	this	morning	not	call	college
where	please	and	car	same	really
your	look	good	road	home	hard
doing	want	face	complete	come	go
worship	you	my	what	can	time
ready	work	great	love	in	$

To evaluate the efficiency of our proposed continuous-SLR system, 157 sign sentences were recorded from each signer which were comprised of 2 to 6 sign words from these 35 different sign gestures. Each signer has repeated these sign sentences. Thus, a total of 942 (6*157) sign sentences were recorded. The distribution of sign words in sentences is depicted in Fig. 8, where sentences with three sign words have maximum number of occurrences.

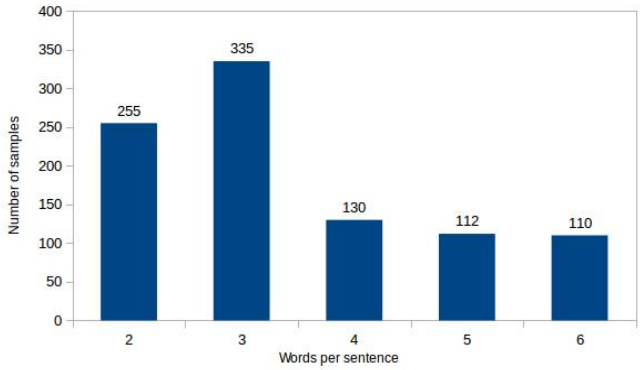

Fig. 8. Number of signed sentence as per the sign words involved in continuous gestures.

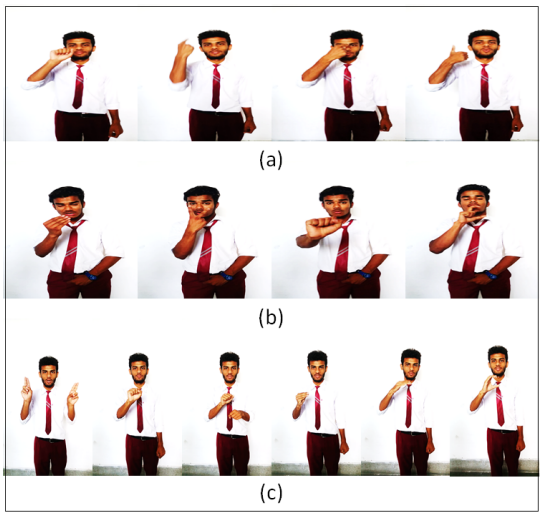

Fig. 9. Few examples of continuous sign sentences: (a) 'your face is good' (b) 'please look your college' (c) 'complete your college and come to college'.

A few examples of different sign sentences are depicted in Fig. 9 which were generated by two different signers. To visualize the variation among continuous sign sequences created by different signers, 3D plots of each hand are depicted in Fig. 10. The plots of the movements recorded in palm center are drawn while performing the sign gestures. It can be observed from the Fig. 10 that there is a similarity in pattern of the same sign sentence. Similarly, variations can be observed in the plots for different signers. The dataset is made available online[1] for future research.

[1] https://sites.google.com/site/iitrcsepradeep7/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2909837, IEEE Sensors Journal

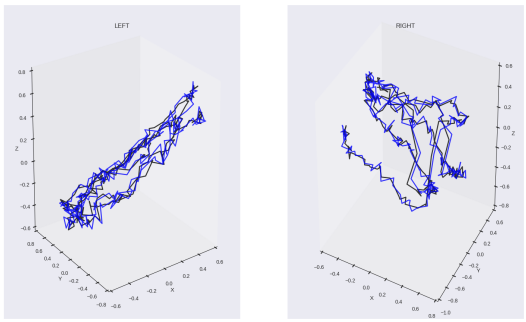

Fig. 10. 3D plots to show the variation among continuous sign words when performed by different signers using both hands for the signed sentence 'doing morning work'.

A. Recognition of Signed Sentences

The recognition of signed sentences was tested by training the proposed modified LSTM model on sign words. The network was trained using categorical cross entropy loss function with 256 hidden state in each LSTM layer, a batch size of 256 and Adaptive learning rate starting from 0.001 reducing by a factor of 0.5 after every 20 epochs. The network was trained for 1000 epochs for 40 hours with a RESET LSTM state condition for transition gestures ($) on a NVIDIA Quadro K620 GPU machine. The learning curve of the model is depicted in Fig. 11 that shows the decrease and saturation in training and validation errors.

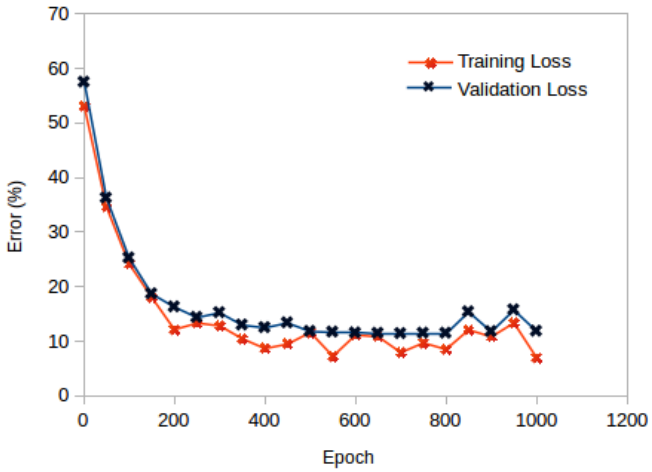

Fig. 11. Learning curve for training and validation errors.

Next, in the testing of signed sentences, an average accuracy of 72.3% was recorded. The variation of accuracies in sign recognition of sentences with different lengths is shown in Fig. 12, where highest recognition rate of 76% was recorded on signed sentences of 2 words whereas the minimum accuracy of 66% was recorded on sentences with six sign words.

The recognition accuracy of sign sentences based on individual signers is depicted in Fig. 13, where maximum accuracy was recorded for signer five which was 77.29%. 5^{th} signer's experience in sign language was highest among all. So, speed and consistency were much better and the gestures were predicted with higher accuracies.

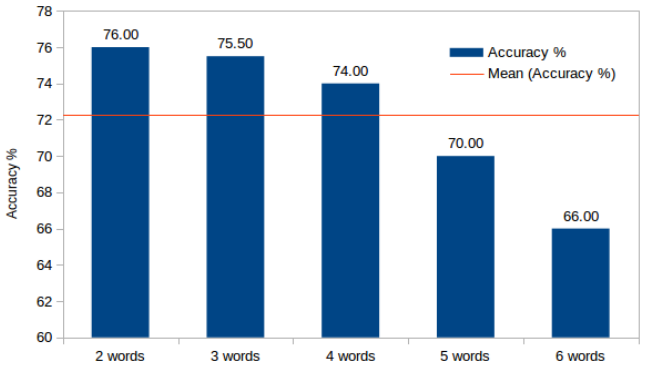

Fig. 12. Recognition accuracies as per the length of signed sentences.

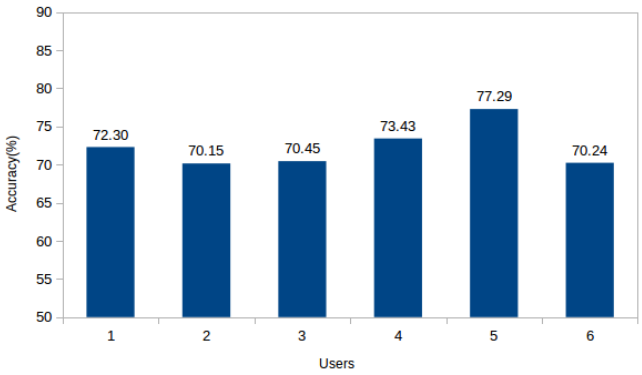

Fig. 13. Recognition accuracies of the signed sentences for each signer.

The recognition process is also carried out by varying different LSTM layers from 2-4 as presented in Table II. It shows that maximum results are recorded with three LSTM layer architecture.

TABLE II
SIGNED SENTENCE RECOGNITION BY VARYING LSTM LAYERS

LSTM Layers	Accuracy (%)
2	59.32
3	**72.3**
4	71.8

In addition, the segmentation process of continuous signed sentences has also been carried out using the proposed architecture. Some of the segmentation results are shown in Fig. 14.

B. Sign Word Recognition performance

Here, we present the recognition results of isolated sign words computed using the proposed modified LSTM model. An average recognition rate of 89.5% has been recorded from 35 isolated sign words. The confusion matrix of all sign words in the form of a heat-map is shown in Fig. 15.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2909837, IEEE Sensors Journal

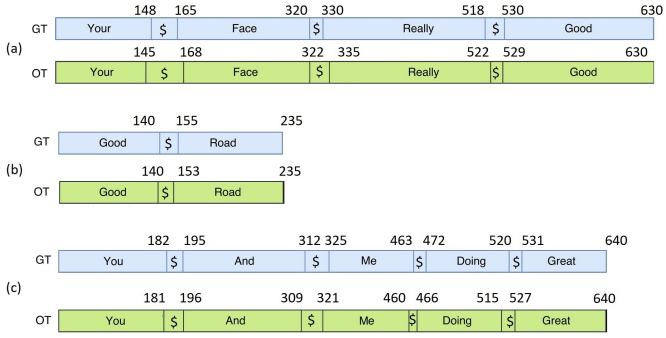

Fig. 14. Results depicting continuous signed segmentation into isolated sign words. GT: Ground truth, OT: output, $: number of frames for transition gesture.

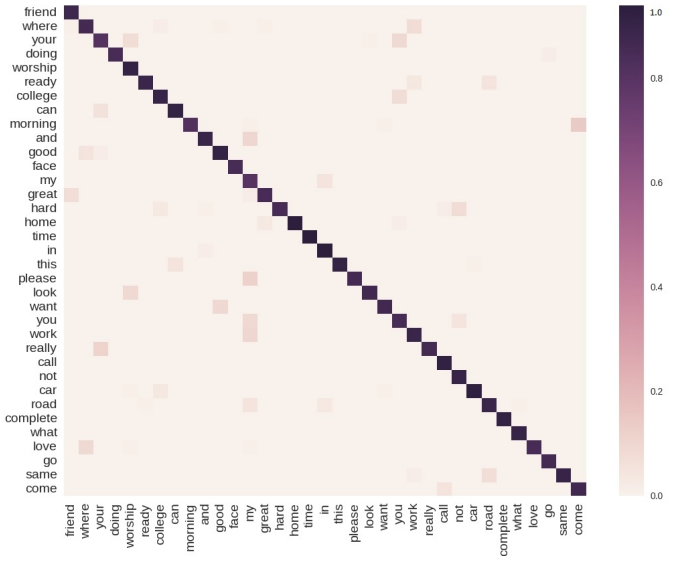

Fig. 15. Heat-map of the confusion matrix for recognition of isolated sign word. Note that on the scale 0 (corresponding color) reflects zero accuracy in the recognition whereas 1 (corresponding color) reflects the maximum recognition rate.

C. Comparative Analysis

In this section, we have performed an indirect comparative analysis of the proposed SLR framework with some state of the art techniques. Kong et al. [9] have proposed continuous SLR system for ASL using the cyber glove. The recognition was performed by a two-layer conditional random field (CRF) model. The lower layer processes the component channels and provides outputs to the upper layer for sign recognition. In their approach, the continuous signed sentences were segmented first before recognition. The methodology was applied on 74 signed sentences with a vocabulary of 107 signs. Similarly, the authors in [1] have proposed the continuous SLR system for Arabic Sign Language (ArSL) using image/video processing techniques. The authors have extracted Discrete Cosine Transform (DCT) based features, and the recognition was performed using HMM classifier. Their approach was tested on 40 sign sentences with a vocabulary of 80 signs. However, our dataset is composed of 157 unique signed sentences. Therefore, we have recorded a recognition performance of 72.3%.

In addition to this, a comparison of the recognition accuracies between the proposed modified-LSTM and a traditional LSTM model is also performed. The training parameters for the traditional LSTMs are 256 hidden states, 256 batch size and Adaptive learning rate starting from 0.001 reducing by a factor of 0.5 after every 20 epochs. The comparison is performed for both signed sentences and the isolated sign words. The recognition accuracies are presented in Table III, where the proposed model outperforms the simple LSTM architecture with a margin of 20.8% and 19.1% in signed words and sentences, respectively. As expected, with CNN's ability to understand the spatial relationship between the data points $f_1 ... f_{12}$ and LSTM's ability to understand the sequence we were able to improve the prediction accuracy.

TABLE III
COMPARATIVE PERFORMANCE ANALYSIS BETWEEN THE PROPOSED AND TRADITIONAL LSTM MODEL.

Model	Sign Word Recognition	Sign Sentence Recognition
Traditional LSTM	68.60 %	53.20 %
Proposed	**89.50 %**	**72.30 %**

V. CONCLUSION

In this paper, we have presented a novel framework for continuous-SLR using Leap motion sensor. A modified LSTM architecture has also been proposed for the recognition of sign words and sentences. A dataset of 35 isolated sign words has been used while training the model. The evaluation of our approach has been performed on 942 signed sentences produced by six signers. Average accuracies of 72.3% and 89.5% have been recorded on the signed sentences and isolated sign words, respectively. In future, the recognition performance can be improved by increasing more training data for better model learning.

REFERENCES

[1] K. Assaleh, T. Shanableh, M. Fanaswala, F. Amin, H. Bajaj, et al. Continuous arabic sign language recognition in user dependent mode. *Journal of Intelligent learning systems and applications*, 2(01):19, 2010. 7

[2] B. Bauer, H. Hienz, and K.-F. Kraiss. Video-based continuous sign language recognition using statistical methods. In *15th International Conference on Pattern Recognition*, volume 2, pages 463–466. IEEE, 2000. 1, 2

[3] B. Bauer and K. Karl-Friedrich. Towards an automatic sign language recognition system using subunits. In *International Gesture Workshop*, pages 64–75. Springer, 2001. 1, 2

[4] S. K. Behera, D. P. Dogra, and P. P. Roy. Analysis of 3d signatures recorded using leap motion sensor. *Multimedia Tools and Applications*, pages 1–26, 2017. 3

[5] H. Cooper, B. Holt, and R. Bowden. Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. Springer, 2011. 1

[6] W. Gao, G. Fang, D. Zhao, and Y. Chen. Transition movement models for large vocabulary continuous sign language recognition. In *International Conference on Automatic Face and Gesture Recognition*, pages 553–558. IEEE, 2004. 2

[7] R. Haldar and D. Mukhopadhyay. Levenshtein distance technique in dictionary lookup methods: An improved approach. *arXiv preprint arXiv:1101.1232*, 2011. 5

[8] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015. 1

[9] W. Kong and S. Ranganath. Towards subject independent continuous sign language recognition: A segment and merge approach. *Pattern Recognition*, 47(3):1294–1308, 2014. 7

[10] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra. Coupled hmm-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, 86:1–8, 2017. 1, 2

[11] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra. A multimodal framework for sensor based sign language recognition. *Neurocomputing*, 2017. 1

[12] P. Kumar, R. Saini, P. Roy, and D. Dogra. Study of text segmentation and recognition using leap motion sensor. *IEEE Sensors Journal*, 2016. 2

[13] P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra. 3d text segmentation and recognition using leap motion. *Multimedia Tools and Applications*, pages 1–20, 2016. 2

[14] Y. LeCun et al. Lenet-5, convolutional neural networks. 3

[15] H. Li and M. Greenspan. Model-based segmentation and recognition of dynamic gestures in continuous video streams. *Pattern Recognition*, 44(8):1614–1628, 2011. 2

[16] K. Li, Z. Zhou, and C.-H. Lee. Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications. *ACM Transactions on Accessible Computing*, 8(2):7, 2016. 1, 2

[17] M. Mohandes, M. Deriche, and J. Liu. Image-based and sensor-based approaches to arabic sign language recognition. *IEEE Transactions on Human-Machine Systems*, 44(4):551–557, 2014. 2

[18] L. Motion. Leap motion controller. *URl: https://www. leapmotion. com*, 2015. 3

[19] L. E. Potter, J. Araullo, and L. Carter. The leap motion controller: a view on sign language. In *25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration*, pages 175–178. ACM, 2013. 1, 2

[20] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos. Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. In *ECCV*, pages 258–272. Springer, 2010. 2

[21] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998. 1, 2

[22] N. Tubaiz, T. Shanableh, and K. Assaleh. Glove-based continuous arabic sign language recognition in user-dependent mode. *IEEE Transactions on Human-Machine Systems*, 45(4):526–533, 2015. 2

[23] C. Vogler and D. Metaxas. Adapting hidden markov models for asl recognition by using three-dimensional computer vision methods. In *International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 156–161. IEEE, 1997. 1

[24] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *Computer Vision and Image Understanding*, 81(3):358–384, 2001. 2

[25] J. Wu, Z. Tian, L. Sun, L. Estevez, and R. Jafari. Real-time american sign language recognition using wrist-worn motion and surface emg sensors. In *12th International Confernece on Wearable and Implantable Body Sensor Networks*, pages 1–6. IEEE, 2015. 2

[26] W. Yang, J. Tao, and Z. Ye. Continuous sign language recognition using level building based on fast hidden markov model. *Pattern Recognition Letters*, 78:28–35, 2016. 2

[27] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. American sign language recognition with the kinect. In *13th international conference on multimodal interfaces*, pages 279–286. ACM, 2011. 1

[28] W. Zhang, K. Itoh, J. Tanida, and Y. Ichioka. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. *Applied optics*, 29(32):4790–4797, 1990. 3

[29] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 1, 2