Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 47 (2026), No. 5-A, pp. 1633–1641
DOI : 10.47974/JIOS-2253

Experimental evaluation of a BiLSTM network for Indian sign language recognition

Silviya Chouhan *
Rajesh Bhadada †
Department of Electronics & Communication Engineering
MBM University
Ratanada
Jodhpur 342011
Rajasthan
India

Abstract

For the Hearing-and speech-impaired (HSI) individuals in India, the ISL functions as a standard and helps communication. This paper proposed an efficient Indian Sign Language recognition system that focused on computational efficiency and accuracy, which is achieved by intelligent temporal modelling, key frame selection and lower-dimension feature extraction. The frames are used to extract features with the Mediapipe Holistic model (MPHm) to form a low-dimensional feature vector. A BiLSTM network, which learns temporal dependencies in both forward and backward directions, is used to model the sequential characteristics of signs. The model has been trained and evaluated on 22 frequently-used ISL isolated words from a public-domain pre-recorded video dataset called INCLUDE and achieved an accuracy of 96.91% on these sample strata. In addition to that, the performance of the model is also evaluated on the INCLUDE 50 dataset, and an accuracy of 98.37% has been achieved.

Subject Classification: *68T07.*

Keywords: *ASLR, Sign language, MPHm, Hearing-and speech-impaired, Indian sign language.*

1. Introduction

In any civilized and developed society, natural communication is a basic human necessity. Humans achieved it mostly orally and like it. However, in society, there are Hearing-and speech-impaired (HSI) people;

* *E-mail:* **silviyachouhan@gmail.com (Corresponding Author)**
† *E-mail:* **rajesh_bhadada@rediffmail.com**

© Taru Publications

it is difficult for them to communicate orally with others as well as amongst themselves. To help them communicate in a standard language, ISL has evolved, enabling communication through gestures rather than spoken words. It empowers HSI individuals to convey their messages using hand movements, facial expressions, and body language, thereby expressing their communication effectively [1]. Due to the lack of understanding of sign language (SL) among the general public and individuals with HSI, effective communication using ISL remains a significant challenge, and an interpreter is necessary [2]. Globally, there are nearly 300 different types of signs [3]. According to the 2011 Census, India had approximately 5 million hearing-impaired and 1.9 million speech-impaired individuals, which is likely to have increased much more by now [4]. In sensor-based techniques, specialised wearable devices are applied to detect and interpret signs [1], [5]. Vision-based methods rely on images (frames) or video (frame sequence) [6].

The rest of this paper is organized as follows: Section II presents a survey of related work on vision-based ASLRS. Section III presents details of the proposed methodology, including the dataset, data augmentation, key frame selection, feature extraction, and the BiLSTM model. Section IV presents the performance analysis and outcomes of the proposed system, along with a comparison to state-of-the-art approaches. Finally, Section V concludes the work and highlights future scope.

2. Related Work

This section gives an overview of recent works on sign language recognition (SLR), focusing on vision-based approaches that are related to word recognition and utilize deep learning. Paper [7] recorded videos of seven words representing the days of the week by extracting image sequences and processing them using ResNet50 for detailed feature extraction. These features were later classified by LSTM and obtained an accuracy of 73.60% on plain backgrounds; however, these features could only obtain 33% accuracy on non-plain backgrounds and hence are not suitable for real-time applications. Paper [8] proposed two advanced ISL translation methods that used Bidirectional LSTM and BERT Transformer on the INCLUDE-50 dataset, employed data augmentation and MediaPipe for feature extraction, and achieved 89.5% accuracy. Their method assumed that key gestures always appear between the 12th and 75th frame, making the approach less sensitive, i.e., the model is likely to misclassify fast or slow signers, thus reducing signer-independence. Paper [9] introduced a

Hybrid CNN-BiLSTM SLR System (HCBSLR) based on 20 words from the INCLUDE dataset. CNN-based was used for automatic feature extraction, and key frames were selected via histogram difference thresholding. VGG-19 and Bilstm were employed for spatial and temporal features extraction, respectively and achieved 87.67% accuracy.

Paper [10] used Inception V3 as a pretrained 48-layer CNN model with 160×120 images, noise removal, and 40° rotation augmentation, followed by layers like Global Average Pooling and *Softmax*. An accuracy of 82.11% was achieved for 20 words using the IIITA-ROBITA dataset. Paper [11] designed a hybrid CNN-LSTM model for Kannada SLR, and the MediaPipe Holistic model (MPHm) was used for feature extraction. CNN was applied for spatial features, and LSTM was used for temporal dependencies, and the classification accuracy obtained was 95%. Lastly, paper [12] used a frame extraction without additional feature processing in a 3D-CNN model with 156,544,980 trainable parameters, achieving 88.24% accuracy.

3. Proposed Methodology

The proposed methodology focused on creating a deep learning-based Indian sign language recognition system (ISLRS) based on pre-recorded videos outlined in Figure 1. The process begins with a video input, whereby the signer is performing signs. To make the model more robust and generalizable, we make use of data augmentation techniques, which generate variations from the original data. The second stage is key frame selection, which only keeps the most informative frames from the video without redundancy, especially at the start and end of the signs. Selected frames are extracted using features, which are numerical

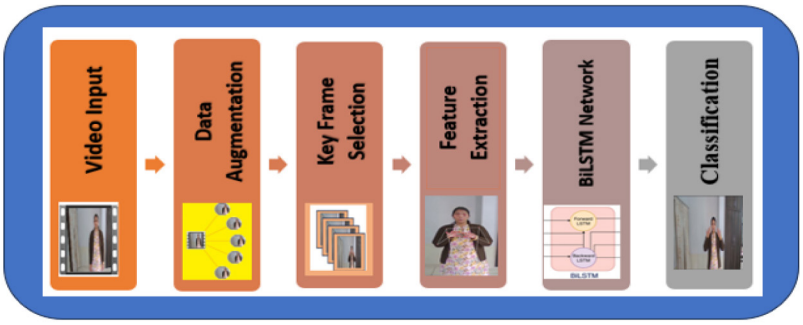


Figure 1

Proposed Methodology

representations of the salient visual information using MPHm. The extracted features are input to a BiLSTM network, which can model the temporal dependency both in forward and backward direction for effectively modelling the dynamic characteristics of signs. At the last stage, a classification layer maps the extracted features to their respective sign labels, enabling effective translation of the gestures into interpretable outputs.

3.1 *Dataset*

For the supply of a standard dataset, the use of public-domain pre-recorded ISL videos, INCLUDE has been selected [13].

3.2 *Data Augmentation*

The proposed method has utilized various data augmentation methods to increase the robustness and diversity of the training dataset while maintaining the dataset's fundamental [13]. To replicate various lighting and colour circumstances, contrast, brightness and saturation vary. The position and size of the gestures within the frame were changed using spatial transformations such as cropping and resizing. In order to balance the dataset, upsampling and downsampling techniques are applied. The original dataset comprised 308 videos for 22 words and 700 videos for the INCLUDE50 dataset, which is significantly expanded to 2,156 and 4900 videos after augmentation for 22 words and INCLUDE50, respectively.

3.3 *Key Frame Selection*

Video data consists of a series of almost identical frames, i.e., successive frames differ marginally from one another. Key frame selection is based on the selection of frames that contain actual sign information, but skip frames that show the signer in a resting position, such as at the beginning and end of a sign. The proposed model uses an algorithm-based automated keyframe selection method to manage the variation in signing speed from different signers, particularly at the start and end of the signs. Compared to traditional methods of temporal segmentation, the proposed keyframe selection method, based on the idea of finding frames where the hands are positioned above the wrist.

3.4 *Feature Extraction*

SL includes complex hand shape, orientation, body posture and facial expression, which makes them characteristically multi-modal. Unlike static signs (single gesture), dynamic signs are made up of continuous motion (sequence of gestures), and become challenging in terms of content recognition because they not only depend on the order but also on the timing of each motion.

3.5 *Model Architecture*

For dynamic sign recognition, future context inclusion reduces the ambiguity of categorization and increases the accuracy of the system by improving the detection of ongoing gestures. To achieve this, a BiLSTM model is used for the classification of sequential data. The proposed architecture processes the input using three stacked Bidirectional LSTMs with 128, 256, and 128 units, respectively, by learning complex temporal dependencies in both forward and backward directions with the *tanh* activation function. To prevent overfitting, dropout layers with a rate of 0.2 have been used after each BiLSTM layer and dense layer. The Adam optimizer has been selected for training the model, the default learning rate of 0.001 and batch size of 32 have been used.

4. Results

To ensure signer diversity across data splits, the dataset is partitioned into 70% for training, 15% for validation, and 15% for testing, which corresponds to the 1508, 324 and 324 numbers of samples, respectively. To thoroughly assess the effectiveness of the model, the performance of the model on the dataset is evaluated using multiple metrics based on a comprehensive evaluation framework, incorporating overall accuracy, precision, recall, and F1-score. Accuracy refers to the degree to which a model's predictions match the true or expected values, and in the case of multiclass classification.

4.1 *Model Performance*

The performance of the proposed Bi-LSTM-based ISL recognition model was evaluated on 22 frequently used words (signs) from the INCLUDE dataset. To support this evaluation, both tabular and graphical representations are employed for a comprehensive comparative assessment. Figure 2 demonstrates the learning behaviour of the model

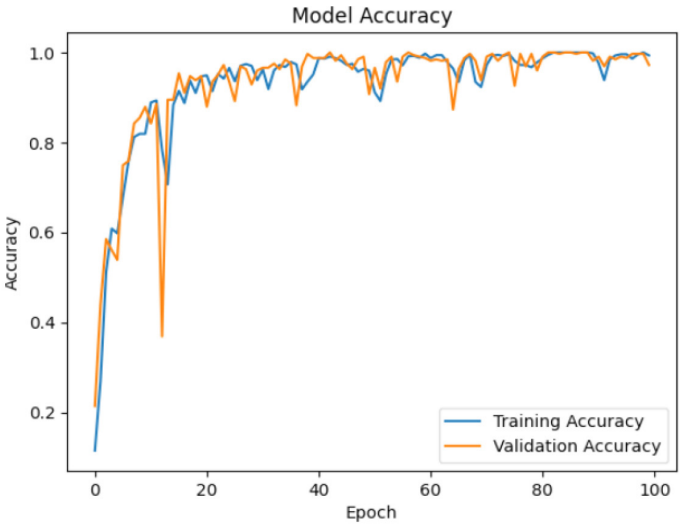


Figure 2

Accuracy vs. Epoch curve

using the Accuracy vs. Epoch curve, where both training and validation accuracy are observed to increase steadily with time.

4.2 *Comparison and analysis*

The proposed model's performance is compared with that of state-of-the-art models. The comparison is based on accuracy and the Inference time. The proposed model achieves the highest classification accuracy of 96.91% against other methods listed in Table 1. The improved performance of the proposed SLR model can be largely attributed to the proper design of the pre-processing pipeline that ensures the input data applied to the model is cleaned, standardized and diversified (augmentation).

These results confirm the superiority of the proposed approach in terms of both efficiency and performance, making it a strong candidate for deployment in real-time ISL recognition systems.

5. Conclusion and Future Scope

This study used the INCLUDE dataset to propose a BiLSTM for the recognition of ISL. To train and evaluate the model, features were extracted from pre-processed video content using the MPHm technique. In this experimental context, the BiLSTM model demonstrated its capacity to

Table 1
Comparison of performance with the state-of-the-art models

Reference	Dataset	Preprocessing	No. of signs	Model	Accuracy
Das, Biswas, and Purkayastha [9]	INCLUDE	Histogram Difference	20	CNN and BiLSTM	87.67%
Baroliya et al. [11]	INCLUDE	Frame Segmentation	20	Hybrid CNN and LSTM	95%
Varsha and Nairy [10]	IIITA-ROBITA	Resize, noise removal, rotation	20	Inception V3 model	82.11%
Singh [12]	Self-Made Hand Gesture	Frame extraction	20	3D-CNN	88.24
Proposed Model	INCLUDE	Data Augmentation, Frame extraction	22	Bi-LSTM	96.91%

capture the temporal dynamics of SL motions with a classification accuracy of 96.91% for 22 frequently used words and 98.37% for the INCLUDE50 dataset. The primary reason for the suggested ASLRS improved performance is that the data applied gets cleaned up, then becomes consistent, and diversified by a preprocessing pipeline, covering data augmentation, key frame extraction, and low-dimensional landmark identification using MPHm. Although the landmarks of MPHm were used as features, misclassification occurred between gestures with similar spatiotemporal patterns (e.g., Happy vs. Sick, Medicine vs. Computer), implying that even detailed features of space might not resolve the problem of gesture ambiguity. In addition, the research utilized a signer-dependent evaluation protocol where data of the same individuals were utilized in training, validation and testing. Signer-independent testing should be used in future to evaluate robustness among users.

References

[1] J. J. David, M. S. S, and S. Revathy, "Indian sign language recogniser with text and speech translation," in *Proc. 8th Int. Conf. Trends Electron. Informatics (ICOEI)*, Tirunelveli, India, pp. 835–840 (2025).

[2] T. Shaikh, T. S. Ansari, A. J. Shaikh, A. A. M. Shaikh, A. N. Khan, and S. S. Katchhi, "Empowering deaf communication: Deep learning framework for sign to speech translation," in *Proc. Int. Conf. Emerging Smart Comput. Informatics (ESCI)*, Pune, India, p. 1–5 (2025).

[3] A. Godbole, R. Dhabliya, V. Deshpande, S. A. Sivakumar, B. M. Shankar, and V. Khetani, "Ethical hacking and penetration testing: Strengthening cybersecurity posture through offensive security measures," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 27, no. 4, pp. 1295–1305 (2024), doi: 10.47974/JDMSC-1983

[4] A. P. Singh and M. Singh, "Real time malware detection in encrypted network traffic using machine learning with time based features," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 26, no. 3, pp. 841–850 (2023), doi: 10.47974/JDMSC-1760.

[5] M. Alaftekin, I. Pacal, and K. Cicek, "Real-time sign language recognition based on YOLO algorithm," *Neural Comput. Appl.*, vol. 36, pp. 7609–7624 (2024).

[6] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep learning for sign language recognition: Current techniques, benchmarks and open issues," *IEEE Access*, vol. 9, pp. 126917–126951 (2021).

[7] P. Mistry, V. Jotaniya, P. Patel, N. Patel, and M. Hasan, "Indian sign language recognition using deep learning," in *Proc. Int. Conf. Artif. Intell. Mach. Vis. (AIMV)*, Gandhinagar, India, pp. 1–6 (2021).

[8] K. B. Prathap, G. D. Swaroop, B. P. Kumar, V. Kamble, and M. Parate, "ISLR: Indian sign language recognition," in *Proc. 2nd Int. Conf. Paradigm Shifts Commun. Embedded Syst., Mach. Learn. Signal Process. (PCEMS)*, Nagpur, India, pp. 1–6 (2023).

[9] S. Das, S. K. Biswas, and B. Purkayastha, "A deep sign language recognition system for Indian sign language," *Neural Comput. Appl.*, vol. 35, pp. 1469–1481, 2023.

[10] M. Varsha and C. S. Nairy, "Indian sign language gesture recognition using deep CNN," in *Proc. 8th Int. Conf. Smart Comput. Commun. (ICSCC)*, Kochi, India, pp. 193–197 (2021).

[11] M. Baroliya, H. Srivastava, A. Iyer, and P. Kannadaguli, "Detection of Indian regional sign language through convolutional neural network," in *Proc. Global Conf. Inf. Technol. Commun. (GCITC)*, Karnataka, India, pp. 1–7 (2023).

[12] D. K. Singh, "3D-CNN based dynamic gesture recognition for Indian sign language modeling," *Procedia Comput. Sci.*, vol. 189, pp. 76–83 (2021).

[13] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, "INCLUDE: A large-scale dataset for Indian sign language recognition," in *Proc. 28th ACM Int. Conf. Multimedia (MM)*, pp. 1366–1375 (2020).

Received April, 2025