Wireless Personal Communications (2023) 132:2527–2541
https://doi.org/10.1007/s11277-023-10730-8

A Spatio-Temporal Framework for Dynamic Indian Sign Language Recognition

Sakshi Sharma[1] · **Sukhwinder Singh**[1]

Accepted: 27 August 2023 / Published online: 13 September 2023

Abstract
A sign language recognition system is a boon to the signer community as it eases the flow of information between the signer and non-signer communities. However, extracting timely detail from the video data is still a challenging task. In this paper, a deep learning based model consisting of trainable CNN and trainable stacked 2 bidirectional long short term memory (S2B-LSTM) has been proposed and tested to recognise the dynamic gestures of Indian sign language (ISL). The CNN architecture has been used as feature extractor to extract the spatial features from the input video data, whereas the temporal relation between the consecutive frames of input video is extracted using S2B-LSTM. This model has been trained and tested on self-developed dataset consisting of 360 videos of ISL dynamic gestures. The CNN-S2B-LSTM model outperforms the existing techniques of sign language recognition with best recognition accuracy of 97.6%.

Keywords Convolutional neural network · Deep learning · Long short-term memory · Indian sign language · Sign language recognition system

1 Introduction

Sign language uses hand gestures and body movement to depict the words of spoken language, allowing hearing-impaired people to communicate with the world [1]. Sign language structure can be decomposed into primary and secondary components that are combined either sequentially or simultaneously. The primary components consist of different hand shapes to make a sign, the location of hands/body, and the movement of hands/fingers/body. In addition to these components, many signs consist of secondary components such as facial expressions and body movements. Among these components, the principal component of sign language is the hand, as most gestures can be represented by it. A gesture is a movement of the hands that creates alphabets, numerals, words, and sentences of the local spoken language. These gestures are categorized into static and dynamic gestures. The static gestures consist of fixed hand positions i.e., no change in position of hands or fingers w.r.t time, whereas dynamic gestures comprise of variable movement of hands[2].

✉ Sakshi Sharma
sak.sharma92@gmail.com

[1] ECE Department, Punjab Engineering College(Deemed to Be University), Chandigarh, India

Sign language syntax differs from region to region, and the language of the Indian community is known as Indian Sign Language (ISL).

The use of sign language is limited to hearing-impaired people as most non-signer people can't understand this language [3]. Sign language recognition system (SLRS) is a useful tool for communicating information between the signer and non-signer community, as it automatically identifies the gestures and translates them into speech or spoken language text [4]. In recent years, the SLRS has created a new way for an interpreter to convert the signs into text/speech. There are numerous successful applications in this field, such as language translation, sign language tutors, and providing special education. These can aid deaf persons in communicating effectively with others. However, due to the intricacy of extracting information from signing components, SLRS remains a difficult process. Learning and constructing a feature vector to represent the information of a hand gesture is challenging as it involves various tasks like tracking hand regions, segmenting hands from the background, discarding irrelevant information, and so on. The approaches presented by different researchers mainly covered static hand gesture recognition. However, the SLRS using merely static gestures, can't handle the large vocabulary and complexity of sign language. Therefore, research into recognizing dynamic gestures is also required to overcome the problem. However, detecting and tracking the complexity of the finger motion activities from the wide-scale body background creates a barrier for dynamic SLR. Another hurdle in dynamic SLR is the extraction of the most discriminative features from the multiple frames of video.

The aim of the proposed work is to recognize the dynamic words of ISL using a vision-based signer-independent system. This proposed method can remember a long–term sequence of two-hand dynamic ISL gestures. The following are the paper's significant contributions:

- An efficient hybridized model with CNN and S2B-LSTM has been designed for dynamic gesture recognition.
- For this, a dataset of 360 videos of dynamic words used by the Indian hearing-impaired community has been created against a uniform background using a camera. Each captured video is converted into a sequence of keyframes only to reduce the computational complexity.
- Experimental findings show a promising result with a recognition accuracy of 97%.

The remaining paper is organized as follows: Sect. 2 contains a detailed report of the existing SLRS. The dataset description is given in Sect. 3. The detail of the proposed method is given in Sect. 4. The experimental evaluation of the work is given in Sect. 5 and finally, Sect. 6 concludes this paper.

2 Related Work

This section aims to briefly discuss the available literature work of SLRs. The available SLRS can be broadly classified into sensor-based [5] and image-based methods [6]. Sensor-based method makes use of external hardware components consisting of different sensors, to capture the signer's detail. The sensors are generally embedded in the gloves, helmet, or body-suit, which signer needs to wear before performing any action. This system is efficient in capturing signing detail, but electronic circuitry restricts the signer's movement.

On the other hand, in vision base method, the signing information can be collected using 2D or 3D camera, which allows the natural movement of hands or body[7]. In the literature, both types of systems has proved helpful for different sign language.

Rekha et al. have presented a recognition system for ISL signs [8]. The authors have collected 23 static and 3 dynamic gestures in this work. Static signs are classified using SVM classifiers and dynamic time warping is used for dynamic signs. Kishore et al. has given a method for recognizing ISL sentences[9]. The collected ISL dataset consisted of a total of 580 sentences. A neural network is used to classify the signs into various words and the performance is computed using a word matching score. Another vision-based approach for ISL recognition is given in[10]. This method has achieved an accuracy of 90% for classifying 24 isolated signs. The use of a leap motion controller for ISL dataset collection has been proposed by Naglot and Kulkarni[11]. ANN is used in this work to classify 26 alphabets and 10 numeric of ISL. Kumar et al. have proposed ASL recognition method using a real-time video [12]. HSV color space model is used for the skin segmentation; the features of static and dynamic gestures are extracted using Zernike moments and curve features, respectively. This work has achieved an accuracy of 93% for static gestures and 100% for dynamic gestures using SVM classifier. Ibrahim et al. presented an automatic recognition method for an Arabic sign language [13]. Firstly, hands are segmented from the dataset of 30 isolated words using the skin-detection technique, and then features like the centre of gravity of hand, motion velocity are used to make feature vector. Then Euclidean distance is used for the classification and it has achieved a classification accuracy of 97%. A convolutional neural-based ASL recognition is presented by Kim et al.[14] The impulse radio sensor is used in this work and the CNN classifier has obtained the classifier accuracy of 90%.

In [15], a vision-based recognition method is presented for ISL static and dynamic gestures. A skin based segmentation is used to extract the signs from the real-time video and then Zernike moments has used to extract key frames from the captured signs. The classification of these signs is done using SVM classifier and it has obtained a recognition rate of 91%. A multimodal ASL recognition system is presented by Ferreira et al.[16]. The colored, depth and leap motion data of 1400 ASL static signs is collected using Kinect and leap sensor. The best recognition accuracy of 97% has been obtained using a CNN classifier. A machine-based interpreter has been designed by Darwish for ArSL [17]. A total of 6000 static samples is collected and classifier using fuzzy HMM.

Another multimodal framework for SLRS is presented by Kumar et al.[18].This method has incorporated facial expression along with the gesture dataset of two sensors. The classification of 51 dynamic gestures is done using HMM. A Vietnamese SLRS is presented in [19]. Microsoft Kinect camera is used to collect the sequence of depth images for 30 dynamic gestures. The performance of the SVM and HMM classifiers is compared in this work and has achieved an average accuracy of 95%. In [20], multimodal dynamic SLRS is presented. The feature extraction and classification of the dynamic gestures is done using 3D convnet and bidirectional LSTM network. This method has achieved a maximum of 89.8% recognition accuracy for Chinese sign language. A spotting-recognition architecture for the recognition of continuous gesture is given in [21].

From the literature review discussed in this section, it can be seen that existing work focuses mainly on static and finger-spelled gestures. Designing a SLRS for dynamic gestures is challenging as it is based on multiple frames. Another point of observation is that a standardized ISL dataset is unavailable. To address these issues, a vision-based dynamic gesture recognition method is presented in this work. The next section gives the detail of the collected ISL dataset.

3 Dataset Creation

Dataset creation is very crucial part of this work. From the literature in the previous section, it is evident that the researcher of every SLRS for ISL have created their own dataset, as there is no publicly available ISL dataset. In this work, a new dataset for dynamic gestures has been collected. The dynamic gestures used for this are from the everyday language of Indian signers. This dataset is collected using camera and it consists of 18 different dynamic gestures. The list of the ISL words used is given in Table 1. In this work, around 4–5 videos has been taken from multiple signers for each dynamic gesture, resulting in total of 360 dynamic videos. To make system versatile, different gender of signer under different lighting conditions are considered. These videos were converted into the sequence of frames and then the keyframes are extracted from the sequence of input frames by removing the frames of negligible hand movement. The pre-processing is done to extract the hand region, and the frames are resized to 256×256 before feeding to the proposed model. The reduced image resolution reduces the computational complexity and speeds up the convergence of classifier. To the best of author's knowledge, this is the first time where a large dynamic ISL dataset has been collected. The samples of this dataset is shown in Fig. 1

4 Proposed Framework

The methodology for the recognition of dynamic gestures and their primary components are discussed in detail in this section. The ISL recognition system is divided into two sections: the feature extraction from the keyframes of input video using 2D-CNN, and then the feature vector is fed to stacked parallel BLSTM for the temporal feature extraction. The framework of the proposed work is also shown in Fig. 2.

4.1 Feature Extraction Using Convolutional Neural Network

CNN is an efficient deep learning method composed of an input layer, convolutional layer, pooling layer, fully connected layer, and output layer [22]. The convolutional network has a wide range of applications. The convolutional layer of this architecture is used as a feature extractor to extract the features automatically from the input feed, and its mathematical equation is given in Eq. 1. It performs the 2D convolution of the input image with the predefined filter.

Table 1 List of ISL dynamic gestures used

ISL dynamic words		
1. Man	7. Tall	13. Television
2. Marry	8. Tap	14. Temple
3. Match	9. Tea	15. Temporary
4. Material	10.Teach	16. Thirty
5. Maximum	11. Teacher	17. Twelve
6. Mean	12. Team	18. Voice

Fig. 1 Sample of extracted keyframes for gesture "Marry"

Several filters with diverse functionalities are utilized to enable the network to extract complementary information and to learn the input characteristic. K different filtered images are produced for the K number of filters used in the convolutional operation. These filtered images are then passed to the pooling layer to reduce the size of the extracted features by using down sampling.

$$f_{xy} = \sum_{i,j} w_{ij} v_{(x+i)(y+j)} + b \tag{1}$$

where f_{xy} is a feature map of position (x, y), w is the weight of kernel, v is the input unit, and b is the bias added to the feature map.

The pooling operation can be done using max-pooling or average pooling. The max-pooling layer splits the input image into a series of the non-overlapping region of a size equivalent to the size of the filter of the pooling layer and then selects the maximum value from each region. However, the average pooling operation selects the average value of the region. By this, the most dominating features of the sub-region are extracted by reducing the spatial size. The pooling operation works independently on every filtered image and resizes them. The CNN models have proved an efficient approach in many practical applications[23], as the image with millions of pixels can be scaled down to the dozens of pixels containing significant characteristics like edges, line, and intensity. This requires the storage of fewer parameters, which further reduces the model's memory requirement and improves its statistical efficiency.

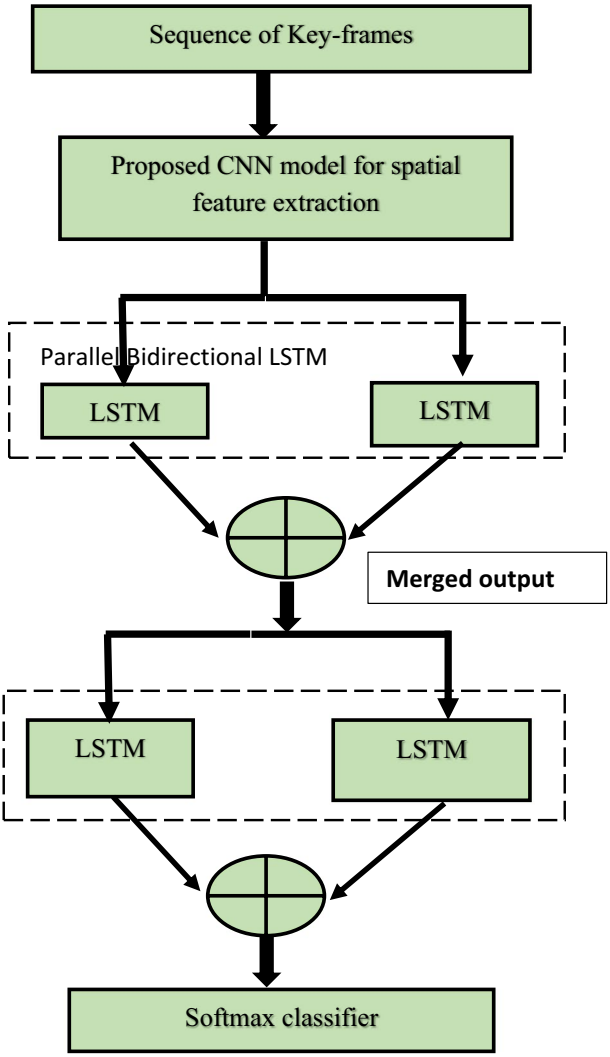

Fig. 2 Framework of the proposed S2B-LSTM

A 2D-CNN model has been built to extract the features from the input videos of dynamic gestures of ISL. A video generally consists of 30 to N frames per second with many redundant frames. In the collected dataset, each video has a total of 110–150 frames, and processing each frame is computationally expensive. Thus, only key frames (generally 13–20 frames per video) are passed to CNN model for feature extraction. It is also evident from the results that 13–20 frames per video don't affect the sign's sequence. The 2D-CNN model of this proposed model consists of three convolutional layers and two pooling layers. This combination of layers has efficiently extracted different global features from the collected dataset and it also gives an advantage of less complex architecture for spatial features extraction. After each layer, a non-linear activation function named RELU is added to introduce some non-linearity in the model. From the experimental analysis of the proposed work, it is clear that this 2D-CNN has successfully extracted all the hidden detail from each frame of gesture. The CNN model can only extract spatial information from the input frames. To learn the relation between the

corresponding frames, a temporal information is required. Thus the extracted features are further fed to S2B-LSTM to learn the change in sequence of the gestures.

4.2 Stacked Bi-directional Long-Short Term Memory (SB-LSTM)

A recurrent neural network is a neural network that uses its internal memory to deal with a sequence of data. The feature handling time series problem has promoted its use in computer vision applications, but regular RNN suffers from the vanishing gradient during the backpropagation process. Hence, RNN models are not capable of learning long-term sequences. The usage of the LSTM network is the solution to this problem. LSTM are a special kind of RNN model proposed by Hochreiter & Schmidhuber [23] to analyse temporal and sequential data [24]. LSTM outperforms RNN network for time series problems as it can deal with short-term and long-term memory requirement problems. The internal structure of LSTM is shown in Fig. 3.

The LSTM approach preserves and saves information's contextual semantics to construct long-term data associations. Its special structure is made up of 4 blocks, which are: cell state, input gate, output gate, and forget gates. LSTM uses these blocks to learn the long-term and short-term sequence. The mathematical operation of LSTM can be given by Eq. 2 to Eq. 7 [26].

$$i_t = \sigma\left(\left(x_t + h_{t-1}\right)W^i + b_i\right) \tag{2}$$

$$f_t = \sigma\left(\left(x_t + h_{t-1}\right)W^f + b_f\right) \tag{3}$$

$$O_t = \sigma\left(\left(x_t + h_{t-1}\right)W^o + b_o\right) \tag{4}$$

$$g = \tanh\left(\left(x_t + h_{t-1}\right)W^g + b_g\right) \tag{5}$$

$$c_t = c_{t-1}.f_t + g.i_t \tag{6}$$

$$h_t = \tanh\left(c_t\right).o_t \tag{7}$$

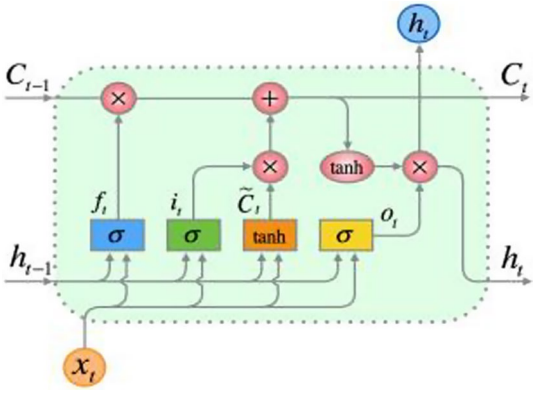

Fig. 3 Internal structure of LSTM [25]

In these equations, x_t is the input at time t, f_t is the forget gate that remembers the previous frame, the information of the upcoming frame is stored by the output gateo_t, g is the recurrent unit, W^i, W^f, W^o, W^g are the weight matrices, σ is the gate activation function. Specifically, the input gate controls and calculates the amount of the current information x_t should be permitted to pass through. The function of forget gate is to ignore the useless information of LSTM past state. The values of forget gate and input gate are adjusted using the sigmoid activation unit during the training phase. The output gate is, also known as recurrent unit with activation function *tanh*, stores the data (o_t) for the next step. The cell state at time t (c_t) is evaluated using the cell state of the previous time frame (c_{t-1}), forget gate value, and the input value of current time stamp and the recurrent unit g. σ refers to the sigmoid function, which gives output in [0,1], tanh is the hyperbolic tangent function that gives output in [-1,1]. At each time iteration (t), the LSTM cell has layer input x_t and layer outputh_t. The complicated networks also takes cell output (c_t) state of time stamp t and cell output (c_{t-1}) of previous time stamp while training and updating the parameters. Due to the gated structure, LSTM handles long-term dependencies, allowing important information to pass through the network. A gated structure enables LSTM to be a useful and scalable model for various sequential data learning applications. It is an effective sequence predictor, particularly for temporal sequence data.

A unidirectional LSTM only retains past information as it reads the input sequence through hidden states, solely in the forward direction. In the case of Bidirectional LSTM (BLSTM), the information is processed in two ways: one is in the forward direction, i.e., from past to future, and another is in the backward direction, i.e., from future to the past, with two separate hidden layers. The final result of the BLSTM is created by combining the outputs of two LSTMSs. For the same sequence of input, BLSTMs yield greater outcomes than LSTMs due to the power of reading in both directions in many fields like speech recognition and phoneme classification. Based on the literature of SLRS, the BLSTM has not been used in ISL recognition.

The structure of unfolded BLSTM model with three consecutive units consisting of a forward layer and backward layer is shown in Fig. 4. The output of both forward and backward layers are computed using the standard equation of LSTM. The output vector $\boldsymbol{y_t}$ of time t is computed by using the Eq. 8.

$$\boldsymbol{y_t} = \sigma(\overrightarrow{h_t}, \overleftarrow{h_t}) \tag{8}$$

where the function σ is used to combine the output of two layers, and this can be either of concatenating function, summation function, multiplication function, or an average function. In this proposed work, a summation function has been used to combine the two output sequences. The final output of the BLSTM (shown in Fig. 4) can be represented by an output vector,$\boldsymbol{Y_T} = [\boldsymbol{y_{t-1}}, \boldsymbol{y_t}, \boldsymbol{y_{t+1}}]$.

In this method, a deep architecture named stacked 2 bi-directional LSTM (S2B-LSTM) neural network, consisting of two BLSTM, has been proposed to learn the long-term dependency of ISL's gestures videos. In this, the output of time t is dependent on previous frames and as well as on upcoming frames. In each BLSTM unit, two LSTM with 24 units are placed in parallel to simultaneously process the input sequence in both forward and backward directions. For the forward pass, the input data is sequentially fed to the model, similar to unidirectional LSTM, whereas the backward component takes the input in reverse order, i.e., from time step to T_x to 1. This proposed model consists of various forward and backward layers to learn the temporal relation between the keyframes of dynamic gestures of ISL. The final output is the concatenation of the output from the

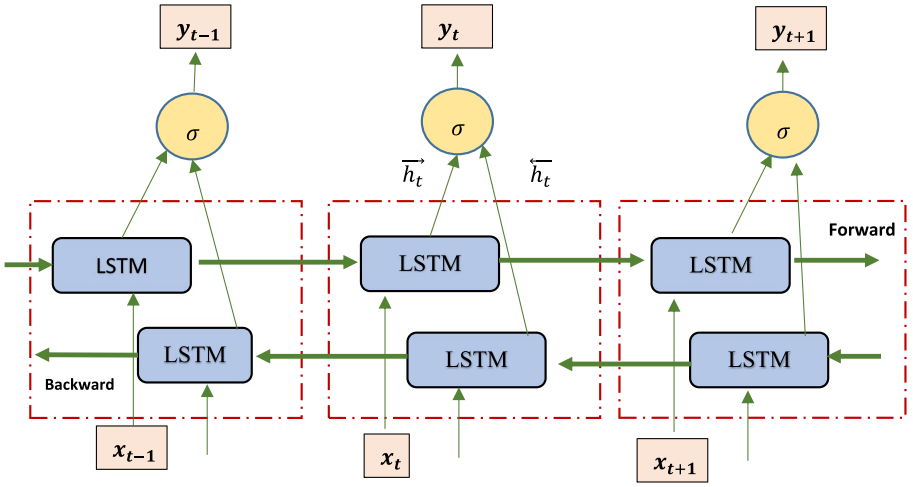

Fig. 4 Bidirectional LSTM with three consecutive units

hidden layers of both LSTMs, which is then fed to the Softmax classifier for the classification. Due to this generative nature of BLSTM, the output layer gets the information from both forward(future) and backward(past) states simultaneously, which effectively improves the context available to the model to learn the relation between the frames. The complete architecture of the proposed model is shown in Fig. 5. It shows that the time series data is fed to the 2D-CNN model for the spatial feature extraction and then these feature vectors are passed to the S2BLSTM model to learn the gestures' time relation and predict the final class.

5 Experimental Evaluation

The detail of the experimental evaluation and the obtained results for the ISL dynamic gestures recognition are discussed in this section. The detail of the dataset used in this paper is already given in Sect. 3. The 80% of this dataset is used for training, and the remaining 20% is used for validation of the method. The training of this model has been carried on Google colab (with GPU runtime). In the training phase, data is fed into the model with small batch size of 128 with a learning rate of 0.001. The other hyper-parameters are also empirically fined-tuned.

5.1 Sign Recognition Results

The experimental analysis of the proposed model is discussed in this section. One of the most commonly used evaluation metrics is "Accuracy, " which is defined as the ratio of correctly predicted classes to the total number of classes used (as given in Eq. 9). The reason for its widespread use is due to the fact that it is simple to calculate, interpret, and summarises the model's capability in a single number.

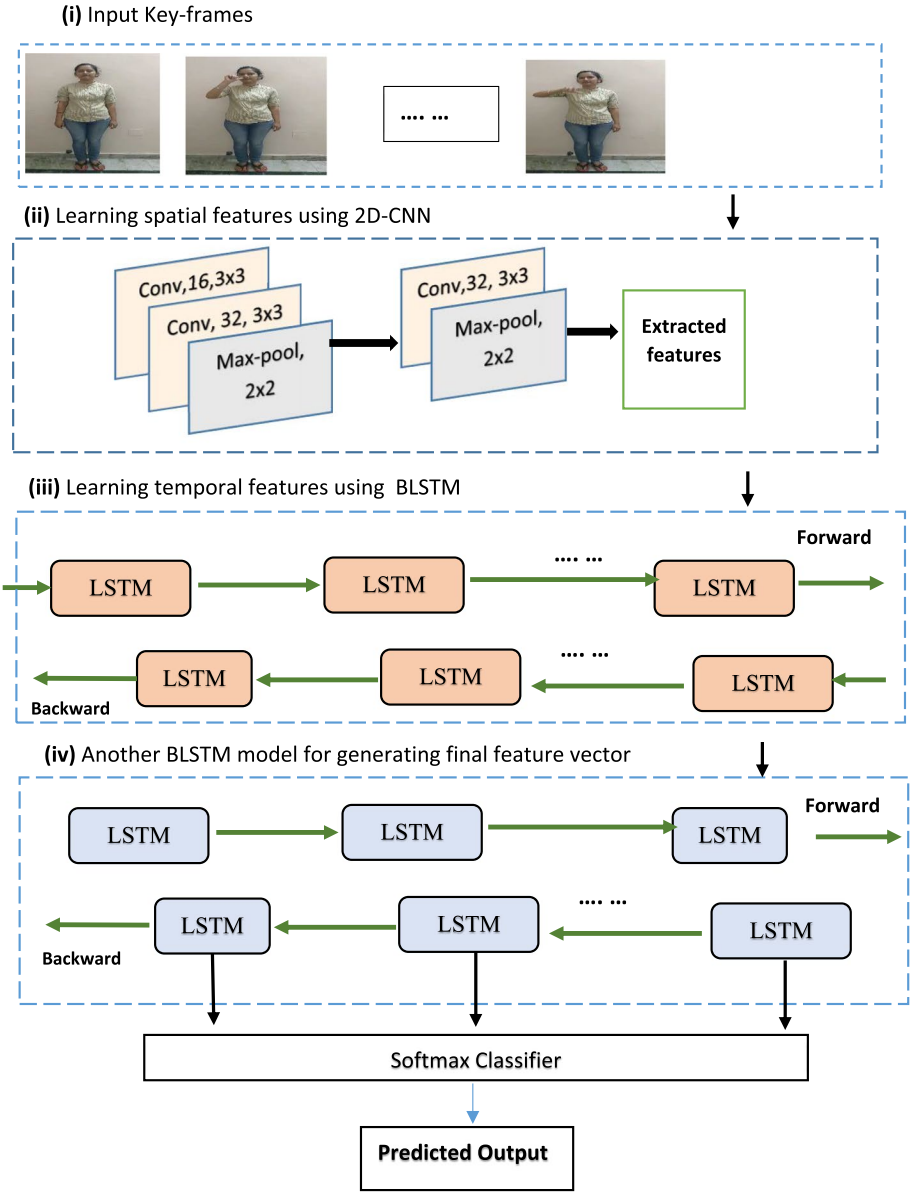

Fig. 5 Architecture of 2D-CNN and SBLSTM model for ISL recognition

$$Accuracy(in\%) = \frac{Correctl\ predicted\ classes}{total\ number\ of\ classes} * 100 \tag{9}$$

For the ISL dataset used, the obtained accuracy plot is shown in Fig. 6. It can be seen that the recognition accuracy increases continuously in the starting iterations and stabilizes at the 40^{th} iteration. This method has achieved maximum recognition accuracy of 97.6%.

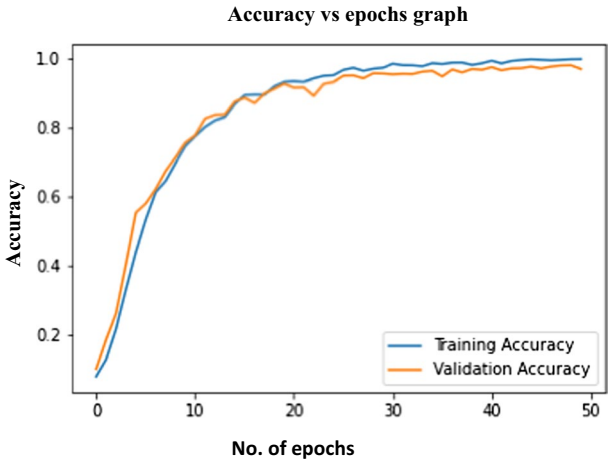

Fig. 6 Accuracy plot for dynamic gestures

As this model has achieved a significant amount of accuracy, hence it can be used in real-world applications to help the deaf community with sign language translation.

The performance of the proposed model is also computed by measuring the value of the computed loss function. It is a way of determining how well a certain algorithm models the data. If the predictions are too far from the actual findings, the loss function will return a large number. On the other hand, it gives a smaller loss function value for a small difference between a predicted and actual result. The loss function learns to reduce the prediction error over time with the help of some optimization functions. A categorical cross-entropy loss function is used to measure the value of the loss while classifying multiple dynamic gestures of ISL. Mathematically this function is expressed using Eq. 10.

$$Loss = \sum_{i=1}^{n} Y_i.log\widehat{Y}_i \tag{10}$$

where, $\widehat{Y}_i$ is the i-th model output value, Y_i is the corresponding target value, and n is the total number of outputs. The plot of loss value obtained for this model is shown in Fig. 7. From this figure, it can be seen that the value of loss drops continuously with the increasing value of iteration. For the dataset of dynamic gestures, the average loss of S2B-LSTM converges to 0.0683.

5.2 Comparison with Other Methods

In this section, the proposed model of S2B-LSTM is compared against the state-of-the-art methods for sign language recognition. Table 2. shows the recognition accuracy of various classifiers on the dynamic gestures. From the literature of SLRS, it is clear that most of the work reported in the literature focused on static gesture recognition, and few research articles are available for ISL dynamic sign recognition. Athira et al.[15] had achieved an accuracy of 89% for 11 dynamic gestures. Rekha et al. [8] tested their model for three dynamic gestures only, and it had achieved an accuracy of 77.2%. Bhuyan et al. [27] tested their model with ten dynamic gestures and obtained an accuracy of 95.8. Ahmed et al. [10] used the DTW method for classifying ISL dynamic gestures and achieved an accuracy of 90%

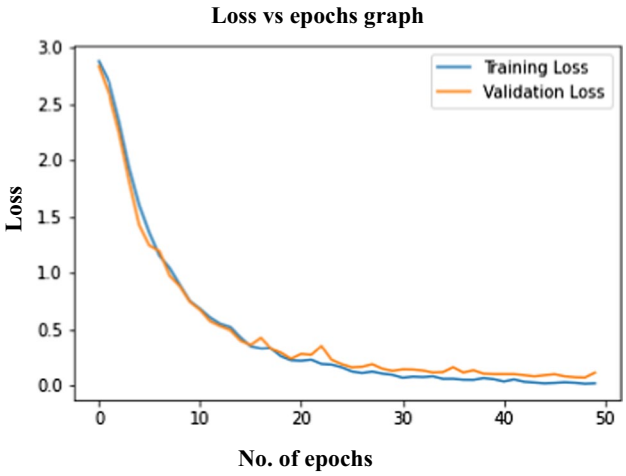

Fig. 7 Loss plot for dynamic gestures

Table 2 Comparison of S2B-LSTM with other methods for ISL dynamic dataset

Methods	No. of gesture	Accuracy (%)
Zernike moments and SVM [15]	26 alphabets (static gestures)	91
	11 dynamic gestures	89
Dynamic time warping [8]	3 dynamic gestures	77.2
Euclidean distance [27]	10 dynamic gestures	95.8
DTW [10]	24 dynamic gestures	90.0
Proposed (CNN + S2B-LSTM)	18 dynamic gestures	97.6

for classifying 24 gestures. Another point of observation from this comparison is that none of the authors had used the temporal sequence learning mechanism for ISL recognition and the efficiency of these mechanisms is unexplored in the field of sign language recognition. The proposed model consisting of 2D-CNN and S2B-LSTM has obtained a maximum accuracy of 97.6%.

6 Conclusion and Future Work

This paper proposed a vision-based approach for the recognition of dynamic gestures of ISL. For this, a hybrid method consisting of CNN and S2B-LSTM is presented. The proposed model of CNN, composed of convolutional and pooling layers, has been used as a feature extractor model to extract the spatial features from the input key frames. Then these feature vector is further passed to S2B-LSTM to extract the video frames' temporal information. The efficiency of this method is tested on the self-collected dataset consisting of 360 videos corresponding to the ISL dynamic gestures. The experimental findings have confirmed the efficiency of the proposed work by yielding a recognition accuracy of 97.6%. Thus the proposed model could efficiently distinguish diverse hand motions by extracting

video spatiotemporal features. There are further area of improvement in the proposed work, like: expanding the size of dataset, testing the performance for thousands of sign gestures, to increase the proposed model's efficiency under unfavourable environment conditions. By incorporating all such ammendments, a vision-based system can also be built to support the communication among hearing-impaired people.

Funding Authors declare that no funding was received for this research work.

Data Availability The authors declare that no data or material was taken illegally. However, a publically available dataset was taken for implementation. The dataset generated in the study is a part of ongoing research work; hence, copyrights are reserved to the institute. Upon completing the ongoing project, this dataset can be made available.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. AL-Rousan, M., Assaleh, K., & Tala'a, A. (2009).Video-based signer-independent Arabic sign language recognition using hidden Markov models. *Appl. Soft Comput.*, *9*(3), 990–999. https://doi.org/10.1016/j.asoc.2009.01.002.
2. Wadhawan, A., & Kumar, P. (2021). Sign language recognition systems: a decade systematic literature review. *Archives of Computational Methods in Engineering, 28*(3), 785–813. https://doi.org/10.1007/s11831-019-09384-2
3. Moreira Almeida, S. G., Guimarães, F. G.& Arturo Ramírez, J. (2014). Feature extraction in Brazilian Sign Language Recognition based on phonological structure and using RGB-D sensors. *Expert Systems with Applications, 41*(16), 7259–7271. https://doi.org/10.1016/j.eswa.2014.05.024.
4. Sharma, S., & Singh, S. (2021). Recognition of indian sign language (ISL) using deep learning model. *Wireless Personal Communications*. https://doi.org/10.1007/s11277-021-09152-1
5. Abhishek, K. S., Qubeley, L. C. F., & Ho, D. (2016). Glove-based hand gesture recognition sign language translator using capacitive touch sensor. In: *2016 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*, Aug. 2016, pp. 334–337. https://doi.org/10.1109/EDSSC.2016.7785276.
6. Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review, 43*(1), 1–54. https://doi.org/10.1007/s10462-012-9356-9
7. S. Sharma and S. Singh, 'Vision-based hand gesture recognition using deep learning for the interpretation of sign language', *Expert Syst. Appl.*, vol. 182, p. 115657, Nov. 2021, doi: https://doi.org/10.1016/j.eswa.2021.115657.
8. Rekha, J., Bhattacharya, J., & Majumder, S. (2011). Shape, texture and local movement hand gesture features for Indian Sign Language recognition. In: *3rd International Conference on Trendz in Information Sciences Computing (TISC2011)*, Dec. 2011, pp. 30–35. https://doi.org/10.1109/TISC.2011.6169079.
9. Kishore, P. V. V., Prasad, M. V. D., Kumar, D. A., & Sastry, A. S. C. S. (2016). Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks. In: *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, Feb. 2016, pp. 346–351. https://doi.org/10.1109/IACC.2016.71.
10. Ahmed, W., Chanda, K., Mitra, S.(2016). Vision based hand gesture recognition using dynamic time warping for indian sign language. In: *2016 International Conference on Information Science (ICIS)*, Aug. 2016, pp. 120–125. https://doi.org/10.1109/INFOSCI.2016.7845312.
11. Naglot, D., & Kulkarni, M. (2016). ANN based Indian Sign Language numerals recognition using the leap motion controller. In: *2016 International Conference on Inventive Computation Technologies (ICICT)*, Aug. 2016, vol. 2, pp. 1–6. https://doi.org/10.1109/INVENTIVE.2016.7824830.

12. Kumar, A., Thankachan, K., & Dominic, M. M. (Mar). Sign language recognition. In: *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, Mar. 2016, pp. 422–428. https://doi.org/10.1109/RAIT.2016.7507939.
13. Ibrahim, N. B., Selim, M. M., & Zayed, H. H. (2018). An Automatic arabic sign language recognition system (ArSLRS). *Journal of King Saud University: Computer and Information Sciences, 30*(4), 470–477, Oct. 2018. https://doi.org/10.1016/j.jksuci.2017.09.007.
14. Kim, S. Y., Han, H. G., Kim, J. W., Lee, S., & Kim, T. W. (2017). A hand gesture recognition sensor using reflected impulses. *IEEE Sensors Journal, 17*(10), 2975–2976. https://doi.org/10.1109/JSEN.2017.2679220
15. Athira, P. K., Sruthi, C. J., & Lijiya, A. (2019). A signer independent sign language recognition with co-articulation elimination from live videos: An indian scenario. *Journal of King Saud University: Computer and Information Sciences.* https://doi.org/10.1016/j.jksuci.2019.05.002.
16. P. M. Ferreira, J. S. Cardoso, and A. Rebelo, 'Multimodal Learning for Sign Language Recognition', in *Pattern Recognition and Image Analysis*, Cham, 2017, pp. 313–321. doi: https://doi.org/10.1007/978-3-319-58838-4_35.
17. Darwish, S. M. (2017). Man-machine interaction system for subject independent sign language recognition. In: *Proceedings of the 9th International Conference on Computer and Automation Engineering—ICCAE '17*, Sydney, Australia, pp. 121–125. https://doi.org/10.1145/3057039.3057040.
18. Kumar, P., Roy, P. P., & Dogra, D. P. (2018). Independent Bayesian classifier combination based sign language recognition using facial expression. *Information Sciences, 428*, 30–48. https://doi.org/10.1016/j.ins.2017.10.046
19. Vo, D.-H., Huynh, H.-H., Doan, P.-M., & Meunier, J. (2017). Dynamic gesture classification for vietnamese sign language recognition. *International Journal of Advanced Computer Science and Applications, 8*(3). https://doi.org/10.14569/IJACSA.2017.080357.
20. Liao, Y., Xiong, P., Min, W., Min, W., & Lu, J. (2019). Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks. *IEEE Access, 7*, 38044–38054. https://doi.org/10.1109/ACCESS.2019.2904749
21. Liu, Z., Chai, X., Liu, Z., & Chen, X. (2017). Continuous Gesture Recognition with Hand-Oriented Spatiotemporal Feature. In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 3056–3064. https://doi.org/10.1109/ICCVW.2017.361.
22. Hagemann, T., & Katsarou, K. (2020). A Systematic review on anomaly detection for cloud computing environments.In: *2020 3rd Artificial Intelligence and Cloud Computing Conference*, New York, NY, USA, Dec. 2020, pp. 83–96. https://doi.org/10.1145/3442536.3442550.
23. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
24. Funahashi, K., & Nakamura, Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks, 6*(6), 801–806.
25. Luo, S., Rao, Y., Chen, J., Wang, H., & Wang, Z. (2020). Short-term load forecasting model of distribution transformer based on CNN and LSTM. *IEEE International Conference on High Voltage Engineering and Application (ICHVE), 2020*, 1–4.
26. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., & Baik, S. W. (2018). Action recognition in video sequences using deep bi-directional LSTM With CNN features. *IEEE Access, 6*, 1155–1166. https://doi.org/10.1109/ACCESS.2017.2778011
27. Bhuyan, M. K., Ghosh, D., & Bora, P. K. (2005). Co-articulation Detection in Hand Gestures. In: *TENCON 2005—2005 IEEE Region 10 Conference*, Nov. 2005, pp. 1–4. https://doi.org/10.1109/TENCON.2005.300947.

Sakshi Sharma is M.E in Electronics and Communication Engineering from Thapar University, Patiala, Punjab, India (now, Thapar Institute of Engineering and Technology (Deemed to be university)). Her area of interest includes soft computing, optimization algorithm, image processing, machine learning, deep learning, pattern recognition, and computer vision. Currently she is pursuing Ph.D. from Punjab Engineering College (Deemed to be University), Chandigarh, India.

Sukhwinder Singh is Ph.D. in Electronics and Communication Engineering from PEC University of Technology, Chandigarh, India (now, Punjab Engineering College (Deemed to be University)). His area of research includes image processing, machine vision and embedded systems. Currently he is working as Assistant Professor in Electronics and Communication Engineering department, Punjab Engineering College (Deemed to be University), Chandigarh, India.