Expert Systems With Applications 185 (2021) 115601

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Applying deep neural networks for the automatic recognition of sign language words: A communication aid to deaf agriculturists

Adithya Venugopalan [*], Rajesh Reghunadhan

Department of Computer Science, Central University of Kerala, Periya, Kasaragod, Kerala, 671320, India

ARTICLE INFO

Keywords:
Sign language recognition
Indian sign language
Dynamic hand gestures
Agricultural words
Complex backgrounds
Deep learning

ABSTRACT

One of the major challenges that deaf people face in modern societal life is communication. For those engaged in agricultural jobs, efficiency at work and productivity are deeply related to the quality of deciphering the sign language used by the deaf farmers. Employing sign language interpreters is not a pragmatic solution to this problem. There comes the need for developing a reliable system for automatic sign language recognition (SLR). This paper reports a work on the recognition of hand gestures for the Indian sign language (ISL) words commonly used by deaf farmers. A hybrid deep learning model with convolutional long short term memory (LSTM) network has been exploited for gesture classification. The model has attained an average classification accuracy of 76.21% on the proposed dataset of ISL words from the agricultural domain.

1. Introduction

Agriculture plays an essential role in the economic development of any country. India, primarily an agricultural land, relies on its vast agricultural workforce in rural areas, where a significant number of workers are completely deaf people, who otherwise are healthy and competent in such work. However, the lack of proper communication facilities is a challenge for deaf individuals working in remote agricultural lands. Even though sign language acts as the voice of deaf people, it fails when they have to interact with the rest of society. This barrier in communication does not allow them to convey their urgent messages (Adithya & Rajesh, 2020b) to the authorities on time. The conventional manual interpretation is not practical in these situations due to its limited availability of interpreters and increased expenditure. Thus, the developments in automatic sign language recognition (SLR) (Ashok et al., 2014; Elakkiya, 2020; Wadhawan & Kumar, 2021) are essential for a better living condition of deaf people, including deaf farmers. Among the various sign language gestures, the developments in recognition of hand gestures (Cheok et al., 2019; Pramod & Martin, 2015; Siddharth & Anupam, 2015) have primary importance as they constitute the letters, digits, words, and phrases of the sign language vocabulary. The COVID-19 pandemic has further increased the role of hand gesture recognition (HGR) in SLR, as communication through other modes like facial expressions and lip reading has been completely blocked by the use of face masks. So HGR for sign language communication has been an important aspect of research in recent years.

The developments in automatic HGR have been started with the instrumented glove interfaces and the electronic sensor-based interfaces that give high accuracy (Divya et al., 2017; Matetelki et al., 2014). However, they are less preferred due to the user inconveniences caused by the complex and relatively expensive hardware setup. Thus, the recent researches on HGR have been moved onto the less expensive, more convenient, and user-friendly approach using visual data (Adaloglou et al., 2021; Cheok et al., 2019; Elakkiya, 2020).

Like any other pattern recognition problem, the success key of vision-based HGR also lies in the robustness and efficiency of the feature descriptors. The state of the art methods for dynamic HGR process the videos of the hand gestures as sequences of images to extract the spatio-temporal feature descriptors through either the classical feature crafting approach (Athira et al., 2019; Bai et al., 2018) or the recently evolved deep learning approach (Adithya & Rajesh, 2020a; Lecun et al., 2015; Mohanty et al., 2017; Traore et al., 2018). This work focuses on the recognition of hand gestures for a set of sign language words that are commonly used by deaf farmers in India. All the words are described by dynamic hand gestures involving both hands.

The current work is aimed to contribute to the developments of SLR in the agricultural domain. A novel video dataset of the hand gestures for the ISL words commonly used by deaf farmers has been proposed. This is one of the few works on ISL recognition with the gesture data collected under uncontrolled conditions like complex backgrounds (including other moving objects), variant illuminations, different camera positions, and orientations. Moreover, this is the first work on SLR

[*] Corresponding author.
E-mail addresses: adithyaushas88@gmail.com (A. Venugopalan), rajeshr@cukerala.ac.in (R. Reghunadhan).

https://doi.org/10.1016/j.eswa.2021.115601
Received 15 September 2020; Received in revised form 6 July 2021; Accepted 11 July 2021
Available online 15 July 2021

that gives a special focus for solving the communication problems of deaf farmers. The work utilizes the recent and powerful deep learning technique through a hybrid deep-net model with the pretrained convolutional neural network (CNN) model, namely GoogleNet (Asifullah et al., 2020) as the front end for feature extraction and the bidirectional LSTM (BiLSTM) sequence network (Sherstinsky, 2020) for feature classification. Although the combination of CNN and LSTM sequence classifier has been used in various video classification tasks, the GoogleNet-BiLSTM pair is applied for the first time in developing an HGR model with the raw gesture videos collected from realistic environments. The classification performance of the deep-net model is validated using a benchmarking hand gesture dataset and has resulted in improved performance compared with the previous results. The model has achieved an average accuracy of 76.21% for the classification of proposed ISL words from the agricultural domain. This result is very promising for dynamic HGR with realistic gesture videos. The reported work will greatly benefit the deaf farmers as it paves a way towards further developments of an automated communication platform for them.

The rest of the paper is organized as follows: Section 2 briefly reviews the previous works on HGR. The framework of the deep classification model for the proposed HGR has been described in Section 3. Section 4 presents the experimental study conducted on the novel hand gesture dataset of ISL agricultural words and Section 5 includes the main conclusions.

2. Literature review

SLR is the task of recognizing sign language gestures and converting them into semantically meaningful words and expressions. As the glosses in the sign language vocabulary are constituted with structured forms of hand gestures, the research in SLR is highly influenced by the research in HGR. There has been significant progress in vision-based HGR using conventional machine learning techniques as well as more powerful deep learning techniques.

Conventional machine learning techniques for HGR involve the extraction of spatio-temporal features from the image sequences of the gestures followed by classification. They follow the classical steps involving image pre-processing, segmentation, hand detection, and tracking before feature extraction. Some of the prominent HGR tasks in this direction include, but are not limited to, the non-linear support vector machine (SVM) classification with the image features and stochastic linear formal grammar(SLFG) (Abid et al., 2015), cyclic pattern estimation of the hand motions through phase alignment (Doan et al., 2017), three level classification scheme with temporal inter frame pattern analysis for static as well as dynamic gestures (Hu et al., 2019), robust part based HGR with three dimensional (3-D) depth features (Ren et al., 2013), SVM model with combined shape and trajectory information (Bai et al., 2018), local binary pattern features with hidden markov model (HMM) (Ahmed et al., 2016), artificial neural network (ANN) model with the discrete cosine transform (DCT) features extracted from the selfie video sequences (Rao et al., 2018), multiclass SVM model trained with hand shape and trajectory features (Athira et al., 2019), hidden conditional neural field (HCNF) classifier with the finger features extracted from the 3-D gesture data captured through leap motion controller (LMC) (Lu et al., 2016), N-dimensional dynamic time warping (ND-DTW) model with 3-D gesture features for human robot interaction (Zhi et al., 2018), hidden markov classification (HMC) of the 3-D features of the hand and finger movements (Vaitkevicius et al., 2019), the linear SVM classification model using the hand kinematic descriptors (Quentin et al., 2019) and the multi class kernel classifier using multi learning features that optimally fuses the hand silhouettes, different figure positions and hand motions (Tao et al., 2021).

The major disadvantage of the conventional machine learning techniques lies in choosing the appropriate feature descriptors. Feature extraction is not embedded as a part of these classification models, and a long trial and error process is needed to decide which features best describe different classes of gestures (Huang & Yang, 2021; Kowdiki & Khaparde, 2021). Conventional approaches to HGR often fail for realistic applications due to many challenging factors like segmentation and tracking of hands from complex and uncontrolled backgrounds, derivation of discriminative feature descriptors, dimensionality reduction and feature selection (Pramod & Martin, 2015), elimination of movement epenthesis (Elakkiya, 2020; Neena & Geetha, 2020) etc. As the hand gesture vocabulary for dynamic SLR shows drastic variations in the appearances and motion patterns, the feature extraction becomes more and more difficult.

The advancements in deep learning networks could overcome these challenges as they avoid the complex image pre-processing and feature extraction steps (Rastgoo et al., 2021). Deep network architectures like CNN automatically learn the high-level features from the raw images/image sequences by eliminating the challenges involved in conventional feature extraction.

Some of the important works on HGR with deep learning techniques include, but are not limited to, the 3-D CNN model with keyframe extraction (Hoang et al., 2018), 3-D attention based residual network (3D-resnet) model (Dhingra & Kunz, 2019), hand skeleton based CNN-LSTM model for 3-D pose recognition (Juan et al., 2018), the combined two dimensional (2-D) CNN and the 3-D dense convolutional network (DenseNet) model (Zhang et al., 2019), 3-D CNN and LSTM with FSM (finite state machine) context awareness model using the RGB and depth video sequences (Hakim et al., 2019), the recurrent neural network (RNN) model with the angles formed by the finger bones of the human hands as features (Avola et al., 2019), Arabic sign language recognition (ArSLR) with a hybrid deep learning model (Aly & Aly, 2020), the multiple deep learning architectures for hand segmentation, feature representations and recognition (Al-Hammadi et al., 2020), the HGR model of CNN with the image enhancement techniques (Neethu et al., 2020), the ISL recognition model using CNN based deep learning (Wadhawan & Kumar, 2020) and the pipelined deep learning architecture for SLR using the multiview hand skeleton features (Rastgoo et al., 2020).

In addition to the above mentioned methods, recent classification models for HGR apply the deep learning technology on the gesture data captured through other active techniques like depth sensors (Ding et al., 2021; Peng et al., 2021), infrared (IR) sensors (Wang et al., 2020) and LMC sensors (Ameur et al., 2020a, 2020b; Lee et al., 2020). It is evident from the literature that the advancements of deep learning technology have made significant improvements in HGR tasks, even for realistic applications. This burst in progress is highly supported by the availability of sufficient data samples for developing the recognition models. As the sign language vocabulary includes a very large number of gesture categories with fewer interclass variations, it is difficult to get sufficient data samples with proper ethical clearance for developing the SLR models for various application domains. The existing works have utilized publicly available hand gesture datasets for developing the models (Adaloglou et al., 2021; Huang et al., 2018) and the most among them include the videos of the hand gestures captured under some standard conditions. Even though there exist a few hand gesture datasets like the American Sign Language (ASL) dataset (Joze & Koller, 2018) collected under variations of light intensity, background objects, camera positions, and orientations, there is no such dataset available on ISL gestures that include the words from the agricultural domain. This work proposes a video dataset of the hand gestures for the ISL words captured from a real agricultural land by placing the camera at different positions and orientations without imposing any control over the background motions and illuminations. The reported SLR task with the realistic gesture videos will contribute to the improvements in ISL recognition in the agricultural domain.

3. Hybrid deep-net model for the proposed ISL word recognition

A hybrid deep learning network of, pretrained 2-D CNN model (Asifullah et al., 2020) GoogleNet and the BiLSTM sequence classifier as depicted in Fig. 1 has been utilized for the proposed ISL recognition. GoogleNet network (Szegedy et al., 2015) extracts the spatial information from the individual video frames. As dynamic HGR is concerned, the classification depends equally well on the temporal patterns too. The BiLSTM network classifier included in the proposed model is utilized to learn the temporal patterns of hand movements from the feature vector sequences extracted with the GoogleNet network. This combination of the GoogleNet network with the BiLSTM sequence classifier is novel for developing SLR models with realistic gesture videos.

3.1. Extracting spatial information using GoogleNet

CNN has shown exemplary performance in image feature extraction with its ability to exploit spatial or temporal correlations in data. The feed-forward multilayered hierarchical structure of CNN performs multiple transformations on the image using a bank of convolutional kernels to extract the useful features from the locally correlated points. The outputs of the convolutional kernels are assigned to the activation functions to perform non-linear transformations on its input for generating different patterns of activation for different responses to capture the semantic variations in images. The outputs from the non-linear activation functions are further followed by a subsampling operation that summarizes the data to make it invariant to geometrical transformations (Asifullah et al., 2020).

The HGR model proposed in this work utilized the pretrained CNN model namely, GoogleNet network (also known as inception v1) (Szegedy et al., 2015) to extract discriminative features from the video frames by incorporating the idea of inception blocks. In GoogleNet, conventional convolutional layers are replaced with the inception blocks encapsulating the filters with different sizes 1 × 1, 3 × 3, and 5 × 5. Inception blocks in GoogleNet are capable of capturing the local spatial information from the image sequences at different scales including both coarse and fine grain levels. The split, transform and merge operations by GoogleNet further enhance the feature extraction with fewer model parameters. All the convolution operations utilized ReLu (Rectified Linear Unit) as activation function. The specially designed architecture makes the GoogleNet learn even the slight variations present in the same type of images with different resolutions (Asifullah et al., 2020; Szegedy et al., 2015). Moreover, the use of the pretrained CNN model for feature extraction eases the training of the proposed HGR model.

The overall architecture of the GoogleNet network is 22 layers deep and it takes RGB images of size 224 × 224 as input (Szegedy et al., 2015). Feature descriptors are the output of the activation function from the last pooling layer "$pool5-7\times7\_s1$" of the GoogleNet network. Each feature vector sequence is an $M \times N$ array, where M is 1024, the size of the feature vector, and N is the number of frames in a gesture video. Fig. 2 shows the frame sequences of the hand gesture videos corresponding to the sign language words "help" and "fire", and their feature vectors are having the sizes 1024 × 45 and 1024 × 56 respectively.

3.2. Learning temporal information with LSTM sequence network

LSTM is a kind of recurrent neural network (RNN) that compensates for the gradient disappearance and gradient explosion of standard RNN during training. It can well classify the long term sequential information with its chain like repeating modules containing three gate controlled cell states. The three gates, namely the forget gate, the input gate, and the output gate, control the information flow through each cell state (Hepeng et al., 2020). The basic structure of an LSTM cell is shown in Fig. 3.

Forget gate takes the output h_{t-1} from the previous cell and the current input X_t at time t and combines them to get the output f_t as given by Eq. (1), where σ is the sigmoid function, w_f and b_f are the trainable parameters that represent the weight matrix and bias values respectively. The output f_t of the forget gate takes the values between 0 and 1 for each cell state (Hepeng et al., 2020). The information from previous state is completely forgotten for $f_t = 0$ and it is passed unaltered for $f_t = 1$.

$$f_t = \sigma\left(w_f[h_{t-1}, X_t] + b_f\right) \tag{1}$$

The input gate decides and stores the information from the new input state X_t through a two-step process involving a sigmoid layer and a tanh layer. Sigmoid layer is the input gate layer i_t that determines which values are going to be updated. It takes the values 0 or 1 calculated as in Eq. (2), where w_i and b_i are the trainable parameters of weight matrix and bias values respectively.

$$i_t = \sigma\left(w_i[h_{t-1}, X_t] + b_i\right) \tag{2}$$

The tanh layer applies a hyperbolic tangent function (tanh) to the current input X_t and the previous output h_{t-1}, returning a vector of new candidate weights $\tilde{C}$. The candidate vector $\tilde{C}_t$ at the tth cell state is calculated as in Eq. (3), where w_C and b_C are the trainable parameters that represent the weight matrix and bias values respectively.

$$\tilde{C}_t = tanh\left(w_C[h_{t-1}, X_t] + b_C\right) \tag{3}$$

$\tilde{C}$ can be added to the internal state of the cell to update its value. Eq. (4) represents the new status of the tth cell state which is updated based on the values of i_t and $\tilde{C}_t$ respectively.

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{4}$$

Finally, the output gate controls how much of the internal state is passed to the output. Eq. (5) gives the output of the sigmoid layer that decides what part of cell state is going to the output. The internal state of the cell is then passed through a tanh layer and multiplied it by the output of the sigmoid layer as in Eq. (6) to get the desired part as the output, where w_o and b_o are the trainable parameters of weight matrix and bias values at the output gate (Hepeng et al., 2020).

$$o_t = \sigma\left(w_o[h_{t-1}, X_t] + b_o\right) \tag{5}$$

$$h_t = o_t \times tanh(C_t) \tag{6}$$

The proposed work utilizes a BiLSTM network model to get enhanced performance for the gesture classification. BiLSTM trains two LSTMs, one on the input sequence as it is and the other on the reverse copy of it. The schematic representation of the BiLSTM network is shown in Fig. 4. The additional context of reverse input sequence in the BiLSTM provides better results for gesture classification using the sequential image features extracted through GoogleNet network.

The architecture of the LSTM sequence network for the proposed hand gesture classification is defined with a sequence input layer with the number of neurons equal to the dimension of the feature vectors, a BiLSTM layer with 2000 hidden units, with each unit containing multiple memory cells, a dropout layer with dropout probability of 0.5, a fully connected layer with the number of neurons corresponding to the number of gesture classes, a softmax layer and a final classification layer that classifies the sequences of feature vectors into the corresponding gesture classes.

3.3. Validating the hybrid deep-net model on a benchmarking dataset

This is the first reported SLR task on the gestures related to agricultural activities and the dataset of ISL agricultural words considered in this work is novel. So the hybrid deep-net model has been first

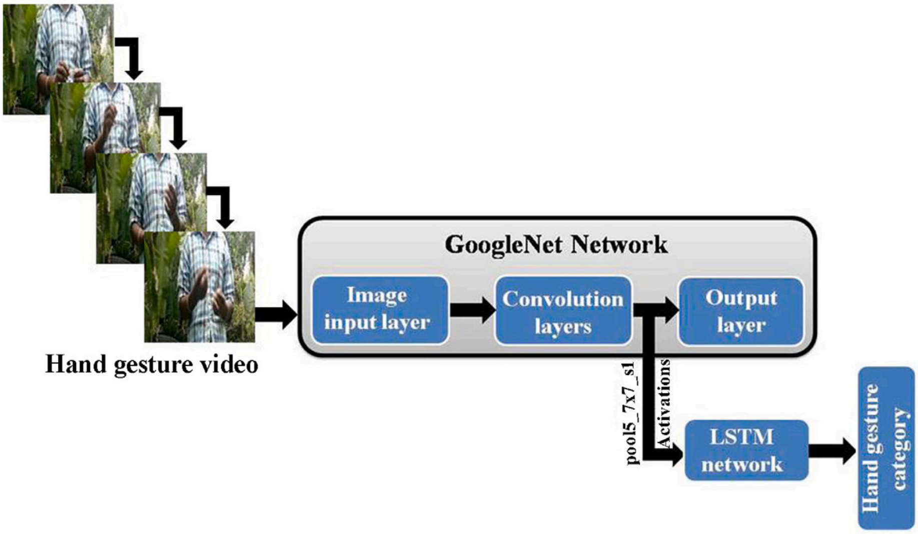


Fig. 1. The architecture of the hybrid GoogleNet-BiLSTM model for the proposed hand gesture classification.

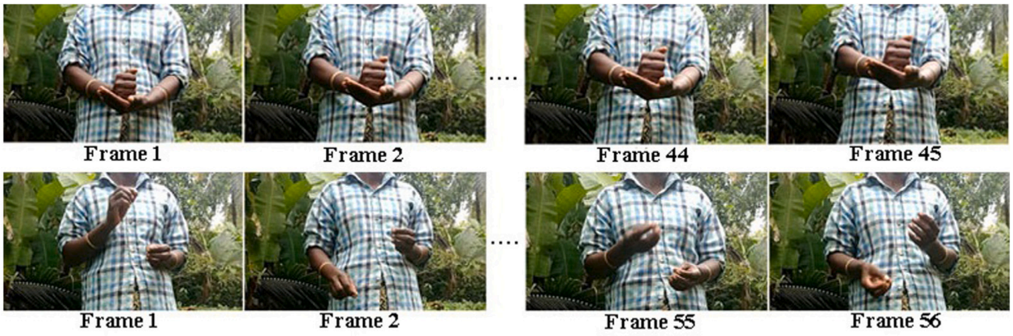


Fig. 2. The first row shows the frame sequences of the hand gesture video correspond to the ISL word "help" and its feature vector size is 1024 × 45, and the second row shows the frame sequences of the hand gesture video correspond to the ISL word "fire" and its feature vector size is 1024 × 56.

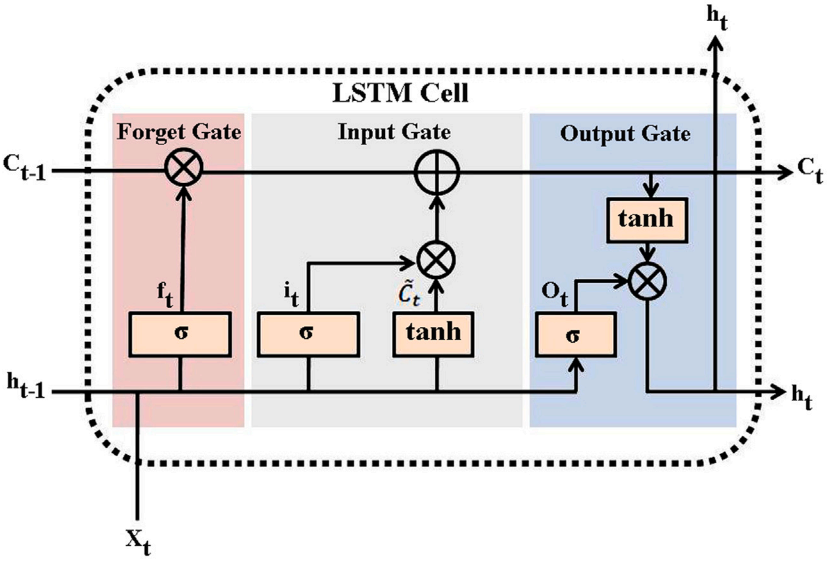


Fig. 3. The structure of an LSTM memory cell with the basic components in a cell state.

applied on the benchmarking Cambridge hand gesture dataset (Kim et al., 2007) to validate its classification performance. The dataset includes 900 video samples of nine hand gesture classes (100 samples for each) defined by three different shapes (flat, spread, and V) and three different motions (left, right, and contract). The video clips were captured under five different illuminations, with 10 arbitrary motions from two individuals. The dataset has been divided into two halves with

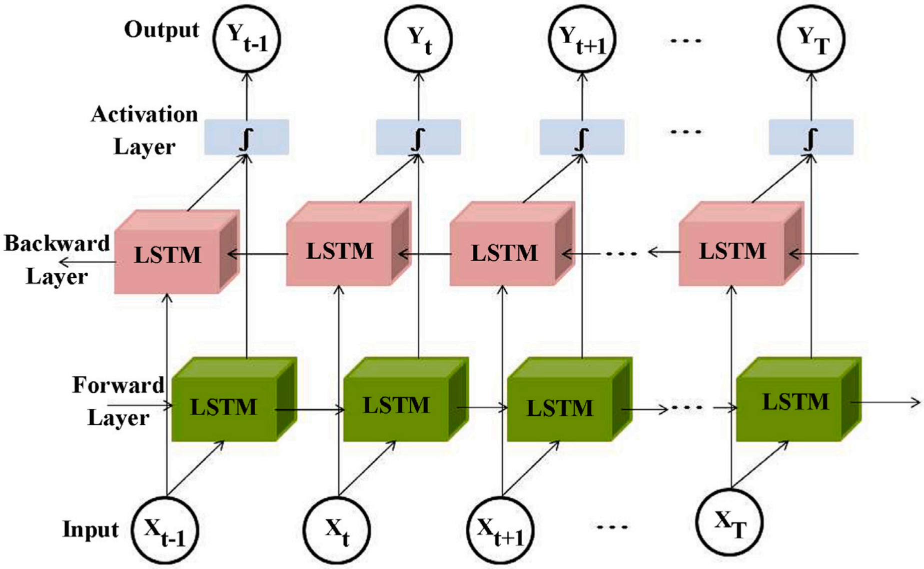


Fig. 4. The basic architecture of a bidirectional LSTM network.

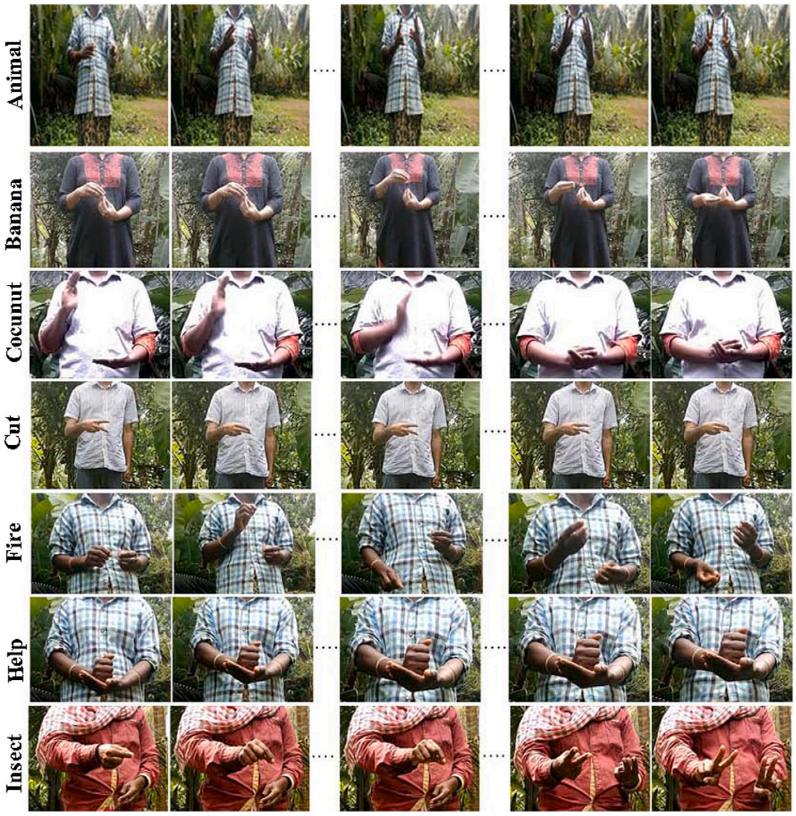


Fig. 5. The frame sequences of the hand gesture videos corresponding to the ISL words “animal”, “banana”, “coconut”, “cut”, “fire”, “help” and “insect” respectively.

450 sample videos for training and the remaining 450 sample videos for testing.

Features are extracted from the raw video sequences using the pretrained GoogleNet model and classified with the BiLSTM network that learns the temporal variations from the feature vector sequences corresponding to each gesture category. The experiment has been carried out with two-fold cross-validation and obtained an average classification accuracy of 97.22 ± 1%. The comparative analysis of the classification accuracy with the results from previous works on the same dataset is given in Table 1. The analysis shows a better performance of the proposed HGR model. The GoogleNet model in the proposed architecture automatically extracts the feature descriptors from the raw image sequences, whereas the compared methods in Table 1 have employed additional operations on the image sequences before feature extraction. In this context, even a slight improvement in the classification accuracy by the hybrid deep-net model can contribute much to further developments in HGR research.

Table 1
Comparison of the classification performance of the GoogleNet-BiLSTM model with the state of the art methods on Cambridge hand gesture dataset.

Author	Method	Accuracy (%)
Kim et al. (2007)	Tensor Canonical correlation analysis	82
Liu and Shao (2013)	Genetic programming using primitive 3-D operators	85
Lui (2012)	Least square regression with tensor model	88
Lui and Beveridge (2011)	Tangent bundle on Grassmann manifold	91
John et al. (2016)	Long term CNN on key video frames	91
Harandi et al. (2013)	3-D covariance descriptors and weighted Riemannian manifold	93
Baraldi et al. (2014)	Dense trajectories hand segmentation	94
Zhao and Elgammal (2008)	Information theoretic keyframe extraction	96.22
Proposed	**GoogleNet-BiLSTM**	97.22 ± 1

4. Experimental study

The hybrid deep learning network explained in Section 3 is used for recognizing the hand gestures for the ISL words from the agricultural domain. As the dataset on the ISL agricultural words is not readily available, a novel video dataset on the same has been created. The detailed experimental study with the description of the proposed dataset is explained in this section.

4.1. Dataset

The lack of publicly available datasets is a major challenge for the development of SLR in many application domains that need emergency communication systems. One such category comes in the agricultural domain and the proposed work focused on the recognition of a set of Indian sign language words used by the deaf farmers for their daily communication. The proposed dataset includes RGB videos of 13 dynamic hand gestures for the words “animal”, “banana”, “coconut”, “cut”, “fire”, “help”, “insect”, “monkey”, “rat”, “rice”, “snake”, “tiger” and “wheat”. The hand gestures were captured according to the styles and movements specified in the ISL dictionary published by Ramakrishna Mission Vivekananda University, Coimbatore, Tamilnadu, India (FDMSE, 2016). This dataset can act as a base for further developments of SLR in this domain.

The videos were captured from four different individuals including two males and two females (two real farmers and two others) in the age group of 25 to 40 years. The data collection was carried out in a real agricultural land on different days in a week, at different instances of time such as early morning, noon, afternoon, and late evening to get the realistic videos for the experiments. The participants were asked to stand comfortably in a land with lots of banana and coconut plantations, and present the hand gestures one by one. The procedure is repeated twice at each instance to capture two sample videos for each gesture. No restrictions have been imposed on the background objects or speed of hand movements so that the dataset contains gesture videos of varying frame lengths which may even include other moving objects like leaves and plants in the background.

The data collection process has got ethical clearance from the Institutional Human Ethics Committee (IHEC) of the Central University of Kerala, Kasaragod, India. All the individuals have gone through the detailed informed consent form and signed their consent for voluntary participation. The dataset contains a total of 260 original videos having the size of 1080 × 1920 pixels with 20 samples for each gesture class. Further data augmentation techniques applied to the gesture videos increase the number of data samples to 932 with 64 sample videos for each gesture class. The frame sequences of the sample videos of all the hand gestures included in the proposed dataset are shown in Figs. 5 and 6 respectively.

4.2. Experimental evaluation

The proposed ISL word recognition has been carried out by dividing the dataset into two halves with 50% of the gesture videos for training and the remaining 50% for testing. In the training phase, the individual video frames were rescaled to 224 × 224 pixels and fed as input to the GoogleNet network for feature extraction. The output of the activation function from the last pooling layer of the GoogleNet network results in 1024 feature values for each frame. The feature vector sequences corresponding to the gesture videos are given as input to the BiLSTM sequence network, in which 90% of the feature vector sequences are used for training and the remaining 10% are used for validating the network. The analysis of the validation results made the BiLSTM network be trained in 20 epochs with an adaptive momentum optimization (adam) function, an initial learning rate of 0.0001, and a gradient threshold of value two, by feeding the data in mini-batches of size 16 to get the maximum recognition rate. Transfer learning by the pretrained GoogleNet model eases the feature extraction process and the model’s training. In the testing phase, the sequences of feature vectors are extracted from the gesture videos using the GoogleNet network and classified into the corresponding gesture categories using the trained BiLSTM network.

The efficiency of the hybrid deep-net classification model on the proposed ISL word dataset has been evaluated in terms of the classification results as well as the computation time. The plot of the confusion matrix for the proposed gesture classification is shown in Fig. 7. The confusion matrix shows the true positive values highest for the words “rice” and “snake”, and lowest for the word “help” in achieving a classification accuracy of 76.21%. On the other hand, the maximum false positive rates are reported for the words “fire” and “insect”, and the highest false-negative rate is reported for the word “help”. It indicates that many of the gesture videos are misclassified into the category of the words “fire” and “insect”. Similarly, a large number of gesture videos for the word “help” are misclassified into the other gesture classes. So to evaluate the performance of the classification model, the analysis must include the measures of the false positives and false negatives of the classification results.

The statistical measures of precision as in Eq. (7), recall as in Eq. (8) and f-score as in Eq. (9) best describe the classification performance in terms of false positives and false negatives, where TP is the true positive value, FP is the false positive value and FN is the false negative value respectively. The precision value indicates the measure of the positive identifications that are actually correct and the recall (sensitivity) value indicates the proportion of the actual positives identified correctly. F-score is measured as the harmonic mean of the precision and recall values and reflects the performance of the classification model. The classification performance of the hybrid deep-net model showing precision, recall, and f-score for each gesture category is shown in Table 2. The analysis shows a greater performance of the model on the proposed ISL gesture dataset collected from a realistic environment. Moreover, the obtained result can act as a benchmark for further improvements of ISL recognition in the agricultural domain.

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

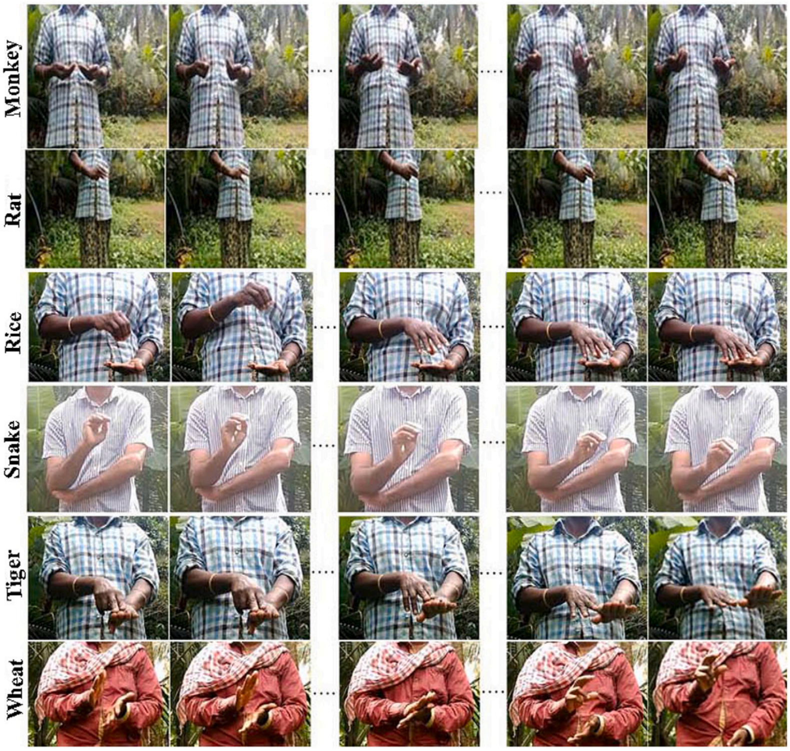
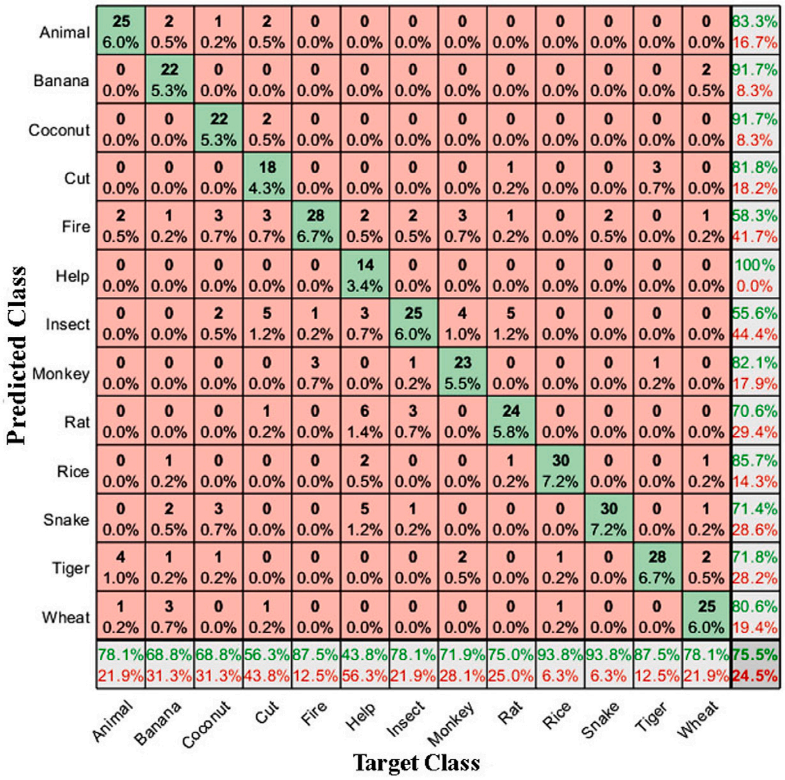


Fig. 6. The frame sequences of the hand gesture videos corresponding to the ISL words “monkey”, “rat”, “rice”, “snake”, “tiger” and “wheat” respectively.



	Animal	Banana	Coconut	Cut	Fire	Help	Insect	Monkey	Rat	Rice	Snake	Tiger	Wheat	
Animal	25 6.0%	2 0.5%	1 0.2%	2 0.5%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	83.3% 16.7%
Banana	0 0.0%	22 5.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.5%	91.7% 8.3%
Coconut	0 0.0%	0 0.0%	22 5.3%	2 0.5%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	91.7% 8.3%
Cut	0 0.0%	0 0.0%	0 0.0%	18 4.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.2%	0 0.0%	0 0.0%	3 0.7%	0 0.0%	81.8% 18.2%
Fire	2 0.5%	1 0.2%	3 0.7%	3 0.7%	28 6.7%	2 0.5%	2 0.5%	3 0.7%	1 0.2%	0 0.0%	2 0.5%	0 0.0%	1 0.2%	58.3% 41.7%
Help	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	14 3.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Insect	0 0.0%	0 0.0%	2 0.5%	5 1.2%	1 0.2%	3 0.7%	25 6.0%	4 1.0%	5 1.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	55.6% 44.4%
Monkey	0 0.0%	0 0.0%	0 0.0%	0 0.0%	3 0.7%	0 0.0%	1 0.2%	23 5.5%	0 0.0%	0 0.0%	0 0.0%	1 0.2%	0 0.0%	82.1% 17.9%
Rat	0 0.0%	0 0.0%	0 0.0%	1 0.2%	0 0.0%	6 1.4%	3 0.7%	0 0.0%	24 5.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	70.6% 29.4%
Rice	0 0.0%	1 0.2%	0 0.0%	0 0.0%	0 0.0%	2 0.5%	0 0.0%	0 0.0%	1 0.2%	30 7.2%	0 0.0%	0 0.0%	1 0.2%	85.7% 14.3%
Snake	0 0.0%	2 0.5%	3 0.7%	0 0.0%	0 0.0%	5 1.2%	1 0.2%	0 0.0%	0 0.0%	0 0.0%	30 7.2%	0 0.0%	1 0.2%	71.4% 28.6%
Tiger	4 1.0%	1 0.2%	1 0.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.5%	0 0.0%	1 0.2%	0 0.0%	28 6.7%	2 0.5%	71.8% 28.2%
Wheat	1 0.2%	3 0.7%	0 0.0%	1 0.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.2%	0 0.0%	0 0.0%	25 6.0%	80.6% 19.4%
	78.1% 21.9%	68.8% 31.3%	68.8% 31.3%	56.3% 43.8%	87.5% 12.5%	43.8% 56.3%	78.1% 21.9%	71.9% 28.1%	75.0% 25.0%	93.8% 6.3%	93.8% 6.3%	87.5% 12.5%	78.1% 21.9%	75.5% 24.5%

Fig. 7. The confusion matrix showing the proposed hand gesture classification results.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

The efficiency of the proposed HGR model has also been evaluated in terms of computation time of classifying one test sequence. The experiment has been conducted on a 2.4 GHz i5-4210 CPU with 8 GB memory, and the proposed HGR is implemented using Matlab. The

Table 2
Classification performance of the hybrid deep-net model on the proposed ISL word dataset showing the precision, recall and f-score values.

ISL Word	Precision (%)	Recall (%)	F-score (%)
Animal	83.33	78.13	80.65
Banana	91.67	68.75	78.57
Coconut	91.67	68.75	78.57
Cut	81.82	56.25	66.67
Fire	58.33	87.50	70.00
Help	100	43.75	60.87
Insect	55.56	78.13	64.94
Monkey	82.14	71.88	76.67
Rat	70.59	75.00	72.73
Rice	85.71	93.75	89.55
Snake	71.43	93.75	81.08
Tiger	71.79	87.50	78.87
Wheat	80.65	78.13	79.37

Table 3
Computation time of each step in the hybrid deep-net HGR model on the proposed ISL word dataset.

Steps	Time
Feature extraction using GoogleNet	5 s to 15 s
Training	7 h
Classification	1 s to 10 s

processing time of each step of the proposed HGR, including feature extraction, training, and classification is estimated, and the values are given in Table 3.

Since the proposed HGR model has been developed with raw videos having unequal numbers of frame sequences, the processing time of each step shows considerable variations among video sequences. The time taken for feature extraction and classification of one test sequence ranges from 5 s to 15 s and 1 s to 10 s, respectively. Even though the feature extraction with pretrained GoogleNet network eases the model's training, it still took approximately seven hours to train the BiLSTM sequence classifier with the whole training data. A possible solution to further reduce the processing time for feature extraction, training and classification is the use of graphical processing unit (GPU) for implementation. The proposed model is scalable and can be applied to even bigger hand gesture datasets with the support of GPU-based implementation.

5. Conclusion

This paper reports deep learning-based dynamic HGR for a set of ISL words commonly used by deaf farmers. A novel dataset of the hand gesture videos for the ISL words has been collected with realistic background and illumination conditions. A hybrid model of GoogleNet network and BiLSTM sequence classifier has been applied for the proposed hand gesture classification and obtained an average accuracy of 76.21%. The proposed work will highly contribute to further developments of ISL recognition in the agricultural domain.

CRediT authorship contribution statement

Adithya Venugopalan: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Rajesh Reghunadhan:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author, Adithya V. thanks Kerala State Council for Science Technology and Environment (KSCSTE), Kerala, India for the research fellowship. The authors express their gratitude to Central University of Kerala, India for the research support. The authors would like to acknowledge all the individuals who have participated in the data collection process. The authors are grateful to Prof. K.J. Thomas, Central University of Kerala, India for helping to revise the manuscript.

References

Abid, M. R., Petriu, E. M., & Amjadian, E. (2015). Dynamic sign language recognition for smart home interactive application using stochastic linear formal grammar. *IEEE Transactions on Instrumentation and Measurement*, *64*(3), 596–605. http://dx.doi.org/10.1109/TIM.2014.2351331.

Adaloglou, N. M., Theocharis, C., Ilias, P., Andreas, S., Th, P. G., Vassia, Z., George, X., Klimis, A., Dimitris, P., & none, D. P. (2021). A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Transactions on Multimedia*, http://dx.doi.org/10.1109/TMM.2021.3070438.

Adithya, V., & Rajesh, R. (2020a). A deep convolutional neural network approach for static hand gesture recognition. *Procedia Computer Science*, *171*, 2353–2361. http://dx.doi.org/10.1016/j.procs.2020.04.255.

Adithya, V., & Rajesh, R. (2020b). Hand gestures for emergency situations: A video dataset based on Indian sign language. *Data in Brief*, *31*, http://dx.doi.org/10.1016/j.dib.2020.106016.

Ahmed, W., Chanda, K., & Mitra, S. (2016). Vision based hand gesture recognition using dynamic time warping for Indian sign language. In *International Conference on Information Science (ICIS)* (pp. 120–125). http://dx.doi.org/10.1109/10.1109/INFOSCI.2016.7845312.

Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., Alrayes, T. S., Mathkour, H., & Mekhtiche, M. A. (2020). Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation. *IEEE Access*, *8*, 192527–192542. http://dx.doi.org/10.1109/ACCESS.2020.3032140.

Aly, S., & Aly, W. (2020). DeepArSLR: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition. *IEEE Access*, *8*, 83199–83212. http://dx.doi.org/10.1109/ACCESS.2020.2990699.

Ameur, S., Khalifa, A. B., & Bouhlel, M. S. (2020a). Chronological pattern indexing: An efficient feature extraction method for hand gesture recognition with leap motion. *Journal of Visual Communication and Image Representation*, *70*, 102842(1–16). http://dx.doi.org/10.1016/j.jvcir.2020.102842.

Ameur, S., Khalifa, A. B., & Bouhlel, M. S. (2020b). A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with leap motion. *Entertainment Computing*, *35*, 100373(1–10). http://dx.doi.org/10.1016/j.entcom.2020.100373.

Ashok, K. S., Gouri, S. M., & Kiran, K. R. (2014). Sign language recognition: State of the art. *ARPN Journal of Engineering and Applied Sciences*, *9*, 116–134.

Asifullah, K., Anabia, S., Umme, Z., & Aqsa, S. Q. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*, 5455–5516. http://dx.doi.org/10.1007/s10462-020-09825-6.

Athira, P., Sruthi, C., & 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*, http://dx.doi.org/10.1016/j.jksuci.2019.05.002.

Avola, D., Bernardi, M., Cinque, L., Foresti, G. L., & Massaroni, C. (2019). Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. *IEEE Transactions on Multimedia*, *21*, 234–245. http://dx.doi.org/10.1109/TMM.2018.2856094.

Bai, X., Li, C., Tian, L., & Song, H. (2018). Dynamic hand gesture recognition based on depth information. In *2018 International Conference on Control, Automation and Information Sciences (ICCAIS)* (pp. 216–221). http://dx.doi.org/10.1109/ICCAIS.2018.8570336.

Baraldi, L., Paci, F., Serra, G., Benini, L., & Cucchiara, R. (2014). Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 702–707). http://dx.doi.org/10.1109/CVPRW.2014.107.

Cheok, M. J., Omar, Z., & Jaward, M. H. (2019). A review of hand gesture and sign language recognition techniques. *Int. J. Mach. Learn. Cyber*, *10*, 131–153. http://dx.doi.org/10.1007/s13042-017-0705-5.

Dhingra, N., & Kunz, A. (2019). Res3atn - deep 3D residual attention network for hand gesture recognition in videos. In *2019 International Conference on 3D Vision (3DV)* (pp. 491–501). http://dx.doi.org/10.1109/3DV.2019.00061.

Ding, W., Ding, C., Li, G., & Liu, K. (2021). Skeleton-based square grid for human action recognition with 3D convolutional neural network. *IEEE Access*, *9*, 54078–54089. http://dx.doi.org/10.1109/ACCESS.2021.3059650.

Divya, B., Delpha, J., & Badrinath, S. (2017). Public speaking words (Indian sign language) recognition using EMG. In *International Conference on Smart Technologies for Smart Nation (SmartTechCon)* (pp. 798–800). http://dx.doi.org/10.1109/SmartTechCon.2017.8358482.

Doan, H., Vu, H., & Tran, T. (2017). Dynamic hand gesture recognition from cyclical hand pattern. In *Fifteenth IAPR International Conference on Machine Vision Applications (MVA)* (pp. 97–100). http://dx.doi.org/10.23919/MVA.2017.7986799.

Elakkiya, R. (2020). Machine learning based sign language recognition: a review and its research frontier. *Journal of Ambient Intelligence and Humanized Computing*, (1–20). http://dx.doi.org/10.1007/s12652-020-02396-y.

FDMSE (2016). *Indian Sign Language (ISL) Dictionary* (third ed.). Faculty of disability management and special education (FDMSE), Ramakrishna Mission Vivekananda University, Coimbatore, India.

Hakim, N. K., Shih, T. K., P, K. S., Aditya, W., Chen, Y. C., & Lin, C. Y. (2019). Dynamic hand gesture recognition using 3Dcnn and LSTM with FSM context-aware model. *Sensors*, *19*(24), 5429(1–19). http://dx.doi.org/10.3390/s19245429.

Harandi, M. T., Sanderson, C., Sanin, A., & Lovell, B. C. (2013). Spatio-temporal covariance descriptors for action and gesture recognition. In *Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV)* (pp. 103–110). IEEE Computer Society, http://dx.doi.org/10.1109/WACV.2013.6475006.

Hepeng, Z., Bin, H., & Guohui, T. (2020). Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recognition Letters*, *131*, 128–134. http://dx.doi.org/10.1016/j.patrec.2019.12.013.

Hoang, N. N., Lee, G.-S., Kim, S.-H., & Yang, H.-J. (2018). A real-time multimodal hand gesture recognition via 3D convolutional neural network and key frame extraction. In *Proceedings of the 2018 International Conference on Machine Learning and Machine Intelligence* (pp. 32–37). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3278312.3278314.

Hu, K., Yin, L., & Wang, T. (2019). Temporal interframe pattern analysis for static and dynamic hand gesture recognition. In *IEEE International Conference on Image Processing (ICIP)* (pp. 3422–3426). http://dx.doi.org/10.1109/ICIP.2019.8803472.

Huang, Y., & Yang, J. (2021). A multi-scale descriptor for real time RGB-D hand gesture recognition. *Pattern Recognition Letters*, [ISSN: 0167-8655] *144*, 97–104. http://dx.doi.org/10.1016/j.patrec.2020.11.011.

Huang, J., Zhou, Q., Li, H., & Li, W. (2018). Video-based sign language recognition without temporal segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence, Vol. 32* (1), (pp. 2257–2264).

John, V., Boyali, A., Mita, S., Imanishi, M., & Sanma, N. (2016). Deep learning-based fast hand gesture recognition using representative frames. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1–8). http://dx.doi.org/10.1109/DICTA.2016.7797030.

Joze, H. R. V., & Koller, O. (2018). Ms-asl: A large-scale data set and benchmark for understanding american sign language. ArXiv Preprint arXiv:1812.01053.

Juan, C. N., Raul, C., Juan, J. P., Antonio, S. M., & Jose, F. V. (2018). Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, *76*, 80–94. http://dx.doi.org/10.1016/j.patcog.2017.10.033.

Kim, T.-K., Wong, S.-F., & Cipolla, R. (2007). Tensor canonical correlation analysis for action classification. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). http://dx.doi.org/10.1109/CVPR.2007.383137.

Kowdiki, M., & Khaparde, A. (2021). Automatic hand gesture recognition using hybrid meta-heuristic-based feature selection and classification with dynamic time warping. *Computer Science Review*, *39*, 100320(1–17). http://dx.doi.org/10.1016/j.cosrev.2020.100320.

Lecun, Y., Bengio, Y., & Lhinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. http://dx.doi.org/10.1038/nature14539.

Lee, A.-r., Cho, Y., Jin, S., & Kim, N. (2020). Enhancement of surgical hand gesture recognition using a capsule network for a contactless interface in the operating room. *Computer Methods and Programs in Biomedicine*, *190*, 105385(1–6). http://dx.doi.org/10.1016/j.cmpb.2020.105385.

Liu, L., & Shao, L. (2013). Synthesis of spatio-temporal descriptors for dynamic hand gesture recognition using genetic programming. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1–7). http://dx.doi.org/10.1109/FG.2013.6553765.

Lu, W., Tong, Z., & Chu, J. (2016). Dynamic hand gesture recognition with leap motion controller. *IEEE Signal Processing Letters*, *23*(9), 1188–1192. http://dx.doi.org/10.1109/LSP.2016.2590470.

Lui, Y. M. (2012). Human gesture recognition on product manifolds. *Journal of Machine Learning Research*, *13*(106), 3297–3321, URL: http://jmlr.org/papers/v13/lui12a.html.

Lui, Y. M., & Beveridge, J. R. (2011). Tangent bundle for human action recognition. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)* (pp. 97–102). http://dx.doi.org/10.1109/FG.2011.5771378.

Matetelki, P., Pataki, M., Turbucz, S., & Kovacs, L. (2014). An assistive interpreter tool using glove-based hand gesture recognition. In *2014 IEEE Canada International Humanitarian Technology Conference - (IHTC)* (pp. 1–5). http://dx.doi.org/10.1109/IHTC.2014.7147529.

Mohanty, A., Rambhatla, S. S., & Sahay, R. R. (2017). Deep gesture: Static hand gesture recognition using CNN. In B. Raman, S. Kumar, P. Roy, & D. Sen (Eds.), *Proceedings of International Conference on Computer Vision and Image Processing. Advances in Intelligent Systems and Computing, Vol 460. Springer*. (460), Springer.

Neena, A., & Geetha, M. (2020). Understanding vision-based continuous sign language recognition. *Multimedia Tools and Applications*, *79*, 22177–22209. http://dx.doi.org/10.1007/s11042-020-08961-z.

Neethu, P., Suguna, R., & Sathish, D. (2020). An efficient method for human hand gesture detection and recognition using deep learning convolutional neural networks. *Soft Computing*, *24*, 15239–15248. http://dx.doi.org/10.1007/s00500-020-04860-5.

Peng, W., Shi, J., & Zhao, G. (2021). Spatial temporal graph deconvolutional network for skeleton-based human action recognition. *IEEE Signal Processing Letters*, *28*, 244–248. http://dx.doi.org/10.1109/LSP.2021.3049691.

Pramod, K. P., & Martin, S. B. (2015). Recent methods and databases in vision based hand gesture recognition: A review. *Computer Vision and Image Understanding*, *141*, 152–165. http://dx.doi.org/10.1016/j.cviu.2015.08.004.

Quentin, D. S., Hazem, W., & Jean, P. V. (2019). Heterogeneous hand gesture recognition using 3D dynamic skeletal data. *Computer Vision and Image Understanding*, *181*, 60–72. http://dx.doi.org/10.1016/j.cviu.2019.01.008.

Rao, G. A., Kishore, P. V. V., Sastry, A. S. C. S., Kumar, D. A., & Kumar, E. K. (2018). Selfie continuous sign language recognition with neural network classifier. In *Proceedings of 2nd International Conference on Micro-Electronics, Electro Magnetics and Telecommunications* (pp. 31–40). Springer.

Rastgoo, R., Kiani, K., & Escalera, S. (2020). Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, *150*, 113336(1–12). http://dx.doi.org/10.1016/j.eswa.2020.113336.

Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, [ISSN: 0957-4174] *164*, 113794(1–27). http://dx.doi.org/10.1016/j.eswa.2020.113794.

Ren, Z., Yuan, J., Meng, J., & Zhang, Z. (2013). Robust part-based hand gesture recognition using kinect sensor. *IEEE Transactions on Multimedia*, *15*(5), 1110–1120. http://dx.doi.org/10.1109/TMM.2013.2246148.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306(1–28). http://dx.doi.org/10.1016/j.physd.2019.132306, URL: https://www.sciencedirect.com/science/article/pii/S0167278919305974.

Siddharth, R., & Anupam, A. (2015). Vision based hand gesture recognition for human computer interaction: A survey. *Artificial Intelligence Review, Springer*, *43*, 1–54. http://dx.doi.org/10.1007/s10462-012-9356-9.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). IEEE, http://dx.doi.org/10.1109/CVPR.2015.7298594.

Tao, S., Honghua, Z., Zhi, L., Hao, L., Yuanyuan, H., & Dianmin, S. (2021). Intelligent human hand gesture recognition by local–global fusing quality-aware features. *Future Generation Computer Systems*, *115*, 298–303. http://dx.doi.org/10.1016/j.future.2020.09.013.

Traore, B. B., Kamsu-Foguem, B., & Tangara, F. (2018). Deep convolution neural network for image recognition. *Ecological Informatics*, [ISSN: 1574-9541] *48*, 257–268. http://dx.doi.org/10.1016/j.ecoinf.2018.10.002.

Vaitkevicius, A., Taroza, M., Blazauskas, T., Damasevicius, R., Maskeliunas, R., & Wozniak, M. (2019). Recognition of American sign language gestures in a virtual reality using leap motion. *Applied Sciences*, *9*, 445(1–16). http://dx.doi.org/10.3390/app9030445.

Wadhawan, A., & Kumar, P. (2020). Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications*, *32*, 7957–7968. http://dx.doi.org/10.1016/j.eswa.2020.113794.

Wadhawan, A., & Kumar, P. (2021). Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering, Springer*, *28*, 785–813. http://dx.doi.org/10.1007/s11831-019-09384-2.

Wang, J., Liu, T., & Wang, X. (2020). Human hand gesture recognition with convolutional neural networks for K-12 double-teachers instruction mode classroom. *Infrared Physics & Technology*, *111*, 103464(1–7). http://dx.doi.org/10.1016/j.infrared.2020.103464.

Zhang, E., Botao, X., Fangzhou, C., Jinghong, D., Guangfeng, L., & Yifei, L. (2019). Fusion of 2D CNN and 3D densenet for dynamic gesture recognition. *Electronics*, *8*, 1511(1–15). http://dx.doi.org/10.3390/electronics8121511.

Zhao, Z., & Elgammal, A. (2008). Information theoretic key frame selection for action recognition. In *Proceedings of the British Machine Vision Conference* (pp. 109(1–10)). BMVA Press, 10.5244/C.22.109.

Zhi, D., de Oliveira, T. E. A., d. Fonseca, V. P., & Petriu, E. M. (2018). Teaching a robot sign language using vision-based hand gesture recognition. In *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)* (pp. 1–6). http://dx.doi.org/10.1109/CIVEMSA.2018.8439952.