Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-022-06843-0

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

Applying Hybrid Deep Neural Network for the Recognition of Sign Language Words Used by the Deaf COVID-19 Patients

Adithya Venugopalan[1] · **Rajesh Reghunadhan**[1]

Received: 13 April 2021 / Accepted: 29 March 2022

Abstract
The rapid spread of the novel corona virus disease (COVID-19) has disrupted the traditional clinical services all over the world. Hospitals and healthcare centers have taken extreme care to minimize the risk of exposure to the virus by restricting the visitors and relatives of the patients. The dramatic changes happened in the healthcare norms have made it hard for the deaf patients to communicate and receive appropriate care. This paper reports a work on automatic sign language recognition that can mitigate the communication barrier between the deaf patients and the healthcare workers in India. Since hand gestures are the most expressive components of a sign language vocabulary, a novel dataset of dynamic hand gestures for the Indian sign language (ISL) words commonly used for emergency communication by deaf COVID-19 positive patients is proposed. A hybrid model of deep convolutional long short-term memory network has been utilized for the recognition of the proposed hand gestures and achieved an average accuracy of 83.36%. The model performance has been further validated on an alternative ISL dataset as well as a benchmarking hand gesture dataset and obtained average accuracies of 97% and $99.34 \pm 0.66\%$, respectively.

Keywords Hand gesture recognition · Indian sign language · COVID-19 · Emergency words · Deep learning

1 Introduction

The evolvement of the COVID-19 pandemic has been severely affecting people from all over the world for the past 2 years. The deaf individuals who make a considerable part of the world population suffer from additional difficulties in this situation due to the existing communication barrier. However, the sign language communication is possible among the deaf community, it fails when they want to interact with the hearing majority of the society [1]. It is not pragmatic to make every person learn and use sign language gestures. These challenges in deaf communication are an ever-existing social concern and have now become more severe for the deaf COVID-19 positive patients.

The pandemic has made the patients stay away from the public including their close relatives for preventing the spread of the virus. The protective guidelines followed by the healthcare centers that keep people safe make communication difficult, broken, and even sometimes impossible for the deaf community. The doctors and staff nurses in the hospitals find it very difficult to understand the sign language gestures used by deaf patients. Many healthcare institutions have taken actions to restrict the visitors and relatives of the patients and to eliminate in-person manual interpreters as a part of social distancing. Even though some hospitals are offering the services of remote interpreters through video conferencing, it often fails due to technical errors. The sudden changes in the healthcare norms have disrupted the clinical services for the deaf and they often find it hard to get appropriate medical care from the health workers. Like any other public, deaf individuals also should be able to report their difficulties and symptoms to bring the current pandemic under control. All these factors have led to the demand for developing an automatic SLR system [2,3] that bridges the new communication barrier between the Deaf and the healthcare workers.

Sign languages are composed of visual gestures formed by hands, face, and other bodily actions, among which the hand gestures form the primary mode of sign language communication. The letters, digits, words, and phrases of the vocabulary are conveyed through hand gestures, while the others play the role of emphasizing their meanings. There are reports on the usage of sign languages from the sixteenth

✉ Adithya Venugopalan
adithyaushas88@gmail.com

[1] Department of Computer Science, Central University of Kerala, Kasaragod, Kerala 671320, India

Published online: 22 April 2022

century onwards. Sign languages have been evolved and used in places where deaf people live. As a result, many variants of sign languages exist in the world like American sign language (ASL), British sign language (BSL), Indian sign language (ISL), Arabic sign language (ArSL), Chinese sign language (CSL), etc. The gesture vocabulary of each sign language was evolved independently based on the regional and cultural variations. The same meaning can be expressed through different forms of gestures in different sign languages. So it is impossible to develop a universal system for SLR.

Currently, the COVID-19 cases are spreading to a lot many people including the deaf. The increase in affected cases and its associated protective measures through social isolation disrupted the healthcare systems for the deaf community. A possible solution for mitigating this issue is to develop an automatic sign language communication system that translates the meanings of sign language gestures into text or voice form to make it understandable by the hearing majority. The most challenging part in developing such an application is the automatic recognition and discrimination of sign language gestures. There exist many video classification works for automatic sign language recognition (SLR) and none of them have achieved compatible recognition accuracy to enable the developments of automatic communication systems for real-life applications. The proposed work focuses to address this challenging issue in the Indian scenario with the recognition of a set of ISL words used by the deaf COVID-19 positive patients for their emergency communication. The work highly contributes to the developments of ISL recognition in the healthcare domain that eases the dissemination of crucial information from deaf patients. A novel dataset of dynamic hand gestures for the ISL words commonly used for COVID-19-related communication is proposed. Gesture videos in the dataset were collected from five different individuals in realistic environments without imposing any restrictions on the background objects and illumination conditions. This is in contrast with majority of the existing works in which the dataset was captured by imposing some restrictions on the backgrounds and illumination conditions.

The proposed SLR utilizes a deep learning model designed as a combination of the convolutional neural network (CNN) [4] and bidirectional LSTM (BiLSTM) [5] sequence network. Although, CNN is a quite common model for image/video-based analysis, the proposed combination of VGG-16 network with bidirectional LSTM is novel for hand gesture classification, and found very optimal on the proposed dataset achieving an average accuracy of 83.36%. The classification performance of the model has also been evaluated on another ISL word dataset as well as a benchmarking dataset, and achieved accuracies of 97% and $99.34 \pm 0.66\%$, respectively. Experimental studies can act as a benchmark for further developments of SLR to break the existing barrier in deaf communication.

The paper is organized as follows. Section 2 presents a detailed report on the existing works on HGR and SLR. The framework of the hybrid convolutional BiLSTM model for the proposed HGR is explained in Sect. 3. Section 4 presents the experimental results and analysis. Finally, Sect. 5 concludes the paper with proper remarks.

2 Related Works

SLR has been a hot topic of research in recent years. As hand gestures are the most structured way of sign language communication, the literature on automatic HGR must also be discussed in this context. HGR has been addressed with electronic sensor-based techniques as well as vision-based techniques [6,7]. Sensor-based techniques measure the finger and hand movement information directly with motion capturing types of equipment. Early works in this direction utilized glove-based methods [8] to acquire the hand gesture data. Then, the works have been gradually evolved into the methods that utilized data captured via radar sensors [9] and electromyogram sensors [10] etc. Even though the sensor-based techniques provide high accuracy, the increased user inconvenience due to the complex and relatively expensive hardware setup makes it the less preferred choice for HGR in realistic applications. More recent works on HGR have been moved on to the vision-based approaches utilizing the images and videos of the gestures [7,11]. The non-invasive and contactless approach to data acquisition has made vision-based techniques the most convenient choice for developing HGR models. Vision-based HGR is implemented through either the traditional pattern recognition approach [6,7] or the more recent deep learning approach [12].

The traditional approach to HGR involves extraction of spatial or spatio-temporal features from image sequences followed by classification. Some of the prominent works in this direction include, but are not limited to, Chinese finger sign language recognition with gray-level co-occurrence matrix features and k-nearest neighbor (KNN) classifier by Jiang et al. [13], HGR model with infrared information captured with the leap motion controller and machine learning techniques like KNN, support vector machine (SVM) and decision trees by Nogales and Benalcazar [14], Grassmann manifold-based discriminant analysis model with the finger tip-based hand trajectory features extracted through either the depth or skeleton information [15], neural network (NN) model with the feature vectors extracted through video summarization technique for Peruvian sign language recognition [16], support vector machine (SVM) classification of the shape and trajectory features extracted via 3-D hand skeleton data [17], SVM model with combined shape and trajectory information by Bai et al. [18], NN model with the textural feature descriptors presented by Agab and Chelali [19], local binary pattern

features with hidden markov model (HMM) for Arabic sign language recognition (ArSLR) by Ahmed et al. [20], artificial neural network (ANN) model with the discrete cosine transform (DCT) features extracted from the selfie video sequences of ISL gestures by Rao et al. [21], and the multiclass SVM model trained with hand shape and trajectory features for ISL words by Athira et al. [22].

The major disadvantage of the traditional HGR approach lies in choosing the appropriate gesture features for each application. 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. As the hand gesture vocabulary for dynamic SLR shows drastic variations in the appearances and motion patterns, the feature extraction becomes more and more cumbersome.

Recent developments in deep learning techniques designed with the concept of automatic feature learning have succeeded in mitigating the challenges of traditional classification models. Deep networks discover the underlying patterns in images and automatically extract out the most descriptive and salient features for each object, as in CASNet proposed by Ji et al. [23] and combined loss-based multiscale fully convolutional network proposed by Li et al. [24]. Some of the notable works on HGR with deep learning architectures include, but are not limited to the dynamic HGR model with multiple deep net architectures for hand segmentation, feature extraction and recognition by Hammadi et al. [25], Arabic sign language recognition (ArSLR) with a hybrid deep learning model by Aly and Walaa [26], sign language word recognition using CNN model with feature fusion approach by Rahim et al. [27], SLR with CNN and hand energy image by Lim et al. [28], combined two-dimensional (2-D) CNN and the 3-D dense convolutional network (DenseNet) model by Zhang et al. [29], 3-D CNN model with keyframe extraction by Hoang et al. [30], 3-D attention-based residual network (3D-resnet) model by Dhingra and Andreas [31], hand skeleton-based CNN-LSTM model for 3-D pose recognition by Juan et al. [32], 3-D CNN, LSTM with FSM (finite state machine) context aware model using RGB and depth video sequences by Hakim et al. [33], CNN model for ArSL recognition by Kamaruzzaman [34], HGR model with CNN architecture by Li et al. [35], deep CNN model for Bangla sign language recognition by Tasmere et al. [36] and the recurrent neural network (RNN) model with the angles formed by the finger bones of the human hands as features by Avola et al. [37].

Despite these promising works, HGR models for ISL recognition still face alot of challenging factors like existence of large vocabulary of gestures that are context-dependent, lack of publicly available datasets for developing the recognition models for various application domains, complicated shapes and movements involving both hands, difference in viewpoints and ways of gesture presentations, segmentation and tracking of hand gestures from uncontrolled backgrounds, derivation of consistent feature descriptors for the huge vocabulary of gestures, temporal variations of gestures, elimination of ambiguous movements between adjacent gestures etc. Moreover, none of the reported works on ISL recognition has addressed the communication barrier between deaf individuals and healthcare workers in the current pandemic situation. The proposed work highlights a way to eliminate the communication difficulties faced by the COVID-19 positive deaf patients in India.

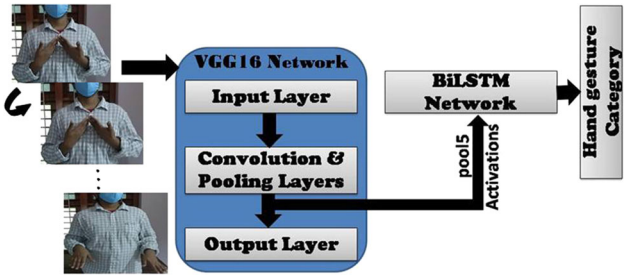

Fig. 1 The architecture of the hybrid deep learning network for the proposed hand gesture classification

3 The CNN-BiLSTM Model for the Proposed ISL Word Recognition

The classification model for the proposed ISL word recognition is shown in Fig. 1. The model is built as a combination of two deep neural network architectures: a CNN-based feature extraction network and an LSTM-based classification network. Videos of the hand gestures correspond to the ISL words are fed as the input to the system. In the feature extraction stage, the gesture videos are converted into sequences of image frames and passed to the CNN model to extract the spatial features from them. Feature extraction and selection have an enormous role in increasing the recognition rate of hand gesture classification. The automatic feature extraction ability of CNN is utilized to complete this task. The feature vectors extracted from individual frames of a gesture video are combined to form a sequence of feature vectors. The feature vector sequences are given as input to the LSTM network for classification. LSTM is a kind of RNN, capable of learning long-term dependencies by handling the vanishing gradient problem.

The proposed approach utilizes the transfer learning technique that allows the reuse of a pretrained deep learning network to extract the image features for training a classifier on top of it. There are classical CNN architectures that have been previously trained on large image datasets and proven their ability to be generalized to images outside them.

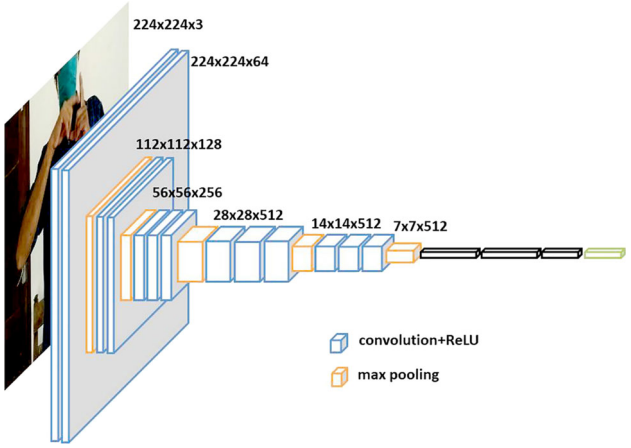

Fig. 2 The VGG-16 network architecture depicting the transformation of a video frame through its layers resulting $7 \times 7 \times 512 = 25088$ feature values from the last max-pooling layer

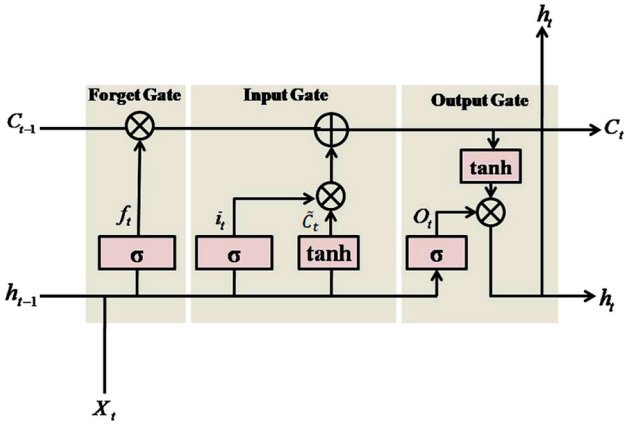

Fig. 3 A typical LSTM memory cell with the basic components of the cell state

The CNN model, VGG-16 [38,39] trained on the ImageNet dataset is retrained in the proposed HGR model to extract the spatial features from the video frames. VGG-16 is a 16 layer CNN model designed with a stacked architecture of convolutional and pooling layers as shown in Fig. 2. All the convolutional layers have the same dimensions for convolution filters as (3,3), and all the max-pooling layers have the same stride size of dimension (2,2). The architecture is composed of two contiguous blocks of two convolutional layers and max-pooling layers, followed by three contiguous blocks of three convolutional layers and max-pooling layers. The stacked convolutional and pooling layers are followed by three fully connected layers with 4096 neurons each and a softmax output layer with 1000 neurons. All the hidden layers are equipped with the rectified linear unit (ReLU) as activation function which enhances the learning process and avoids the vanishing gradient problem. Moreover, the use of small-size filters provides the benefit of low computational cost with fewer hyperparameters for the model.

The individual frames of the gesture videos are passed sequentially to the VGG-16 network after resizing to a dimension of $(224 \times 224 \times 3)$ that matches with the input dimension of VGG-16. The feature values are extracted from the last pooling layer "$pool5''$ of the VGG-16. The resulting feature vector sequences are $M \times N$ matrices, where M is the size (25088) of the feature vectors, and N is the number of frames in a gesture video. The feature vectors extracted from the video frames are grouped to form the sequence feature vectors. Feature extraction using the VGG-16 network derives the most salient representations from the frame sequences of the gesture videos. The use of the pretrained CNN model for feature extraction eases the training of the full classification model.

The feature vector sequences obtained from the VGG-16 network are given as input to a BiLSTM network [5,40] for classification. An LSTM network is a kind of RNN that can model the long-term dependencies in data. Its structure resembles the basic structure of an RNN with repeated chain-like modules called cells. Each LSTM cell contains three gates interacting as shown in Fig. 3.

A typical LSTM cell takes the output of the previous LSTM cell h_{t-1} at time $t-1$ and the current input X_t at time t. The sigmoid function in the forget gate f_t determines which part from the previous output should be eliminated. Equation (1) gives the values of f_t ranging from 0 to 1, where σ is the sigmoid function and W_f and b_f are the weight matrix and bias values, respectively.

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

Next step 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 function determines the new information to be updated i_t giving the values 0 or 1. The tanh function creates a vector of candidate weights $\tilde{C}$ to assign to the values that pass to the next gate and decides their level of importance by assigning the values from -1 to 1. The values of i_t and the candidate vector $\tilde{C}_t$ at the t^{th} cell state are calculated as in Eqs. (2) and (3), respectively, where w_i, b_i, w_c and b_c are the trainable values [41].

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

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

The two values are multiplied to update the new cell state as given in Eq. (4).

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

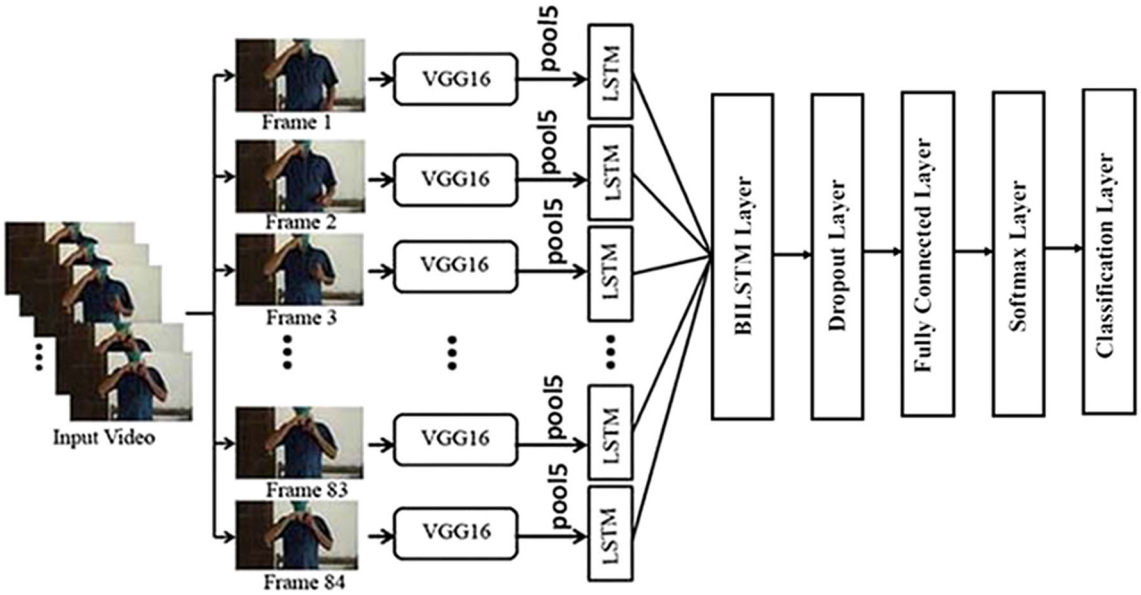

Fig. 4 Block diagram depicting the different steps involved in training the VGG16-BiLSTM network for the proposed hand gesture classification

The output cell state O_t is calculated as in Eq. (5), where the sigmoid layer determines which part of the cell state to be given as the output and, w_o and b_o are the trainable parameters of weight matrix and bias values at the output gate. Finally, the values of h_t are determined as a filtered version of the output cell state O_t as given in Eq. (6). The output of the sigmoid gate O_t is multiplied by the new values created by the tanh layer from the cell state C_t that ranges from (-1 to 1).

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

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

The BiLSTM network architecture utilized in the proposed work has further enhanced the performance of the classifier in learning the temporal information. It trains two LSTMs, one on the input sequence as it is and the other on the reverse copy of it. The overall architecture of the classification network is defined with a sequence input layer with the number of neurons equal to the dimensions of the feature vectors, a BiLSTM layer with 2000 hidden units with multiple memory cells in each unit, a dropout layer with a dropout probability of 0.5, a fully connected layer, a softmax layer, and the final classification layer. Thus, the combination of the VGG-16 network and the BiLSTM sequence network makes a hybrid deep neural network model for the proposed ISL word recognition.

The steps involved in training the proposed hand gesture classification network is depicted with the example of processing a single sample video in Fig.4. The row video of the gesture with 84 frames is passed as the input to the network. Feature descriptors are extracted from each frame of the video using the transfer learning mechanism by VGG-16 network from its last pooling layer ("$pool5''$), and passed to the classification network consisting of the BiLSTM layer, followed by the dropout layer, the fully connected layer, the softmax layer and the final classification layer. Similarly, 680 (50% of the total 1360 samples) gesture videos belonging to different gesture classes are passed to the VGG16-BiLSTM classification network to train the HGR model that can predict different gesture classes in the proposed dataset of ISL words used by COVID-19 patients.

3.1 Validating the Effectiveness of the Proposed VGG16-BiLSTM Model for HGR

HGR has been addressed with a wide variety of methodologies for various applications including SLR and achieved promising results. The work reported in this paper is focused on the recognition of sign language hand gestures used to express the health-related words used by the deaf COVID-19 patients. This is the first work reported on ISL recognition that addresses the communication problem of deaf COVID-19 patients. The proposed work utilizes a novel video dataset of ISL words from healthcare domain that are mainly used to convey the ailments and symptoms of COVID-19. Hence, an explicit comparison of the proposed work with the existing works is not possible. So in order to validate the performance of the proposed VGG16-BiLSTM model, it is first applied on the benchmarking Cambridge hand gesture dataset [42].

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) captured under five different illumina-

Table 1 Performance evaluation of the VGG16-BiLSTM model on the benchmarking Cambridge hand gesture dataset

Author	Method	Accuracy %
Kurmanji and Ghaderi [43]	3D CNN	75.8
John et al. [44]	Long term CNN On key video frames	91
Yui Man Lui [45]	Higher order SVD Grassmann manifold	91.7 ± 2.3
Sanin et al. [46]	3-D covariance descriptors and Weighted Riemannian manifold	93
Baraldi et al. [47]	Dense trajectories Hand segmentation	94
Chandra and Jawahar [48]	Partial Least Squares kernel	94 ± 2.1
Souza et al. [49]	Enhanced Grassmann discriminant analysis Randomized time warping	95.14
Zhao and Elegammal [50]	Information theoretic Keyframe extraction	96.22
Hoang et al. [51]	Keyframe extraction 3D ResNet	97.8
Tang et al. [52]	Keyframe extraction Feature fusion	98.23 ± 0.84
Proposed	**VGG16-BiLSTM**	99.34 ± 0.66

tions from two individuals. The dataset has been divided into two sets with 450 sample videos for training and the remaining 450 sample videos for testing. Features are extracted from raw video sequences using the pretrained VGG-16 network and classified with the BiLSTM network that learns the temporal variations from the feature vector sequences corresponding to each gesture category. The experiment has obtained an average classification accuracy of 99.34±0.66%. A comparison 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 VGG16-BiLSTM model for hand gesture classification.

The proposed deep net model is also utilized to classify an alternative ISL word dataset for emergencies to further evaluate its performance. The dataset includes the videos of ten dynamic hand gestures that correspond to the words "accident," "call," "cut," "doctor," "help," "hot," "lose," "pain," "police," and "thief" (eight among them were published in [1]). Such a dataset is very beneficial to the deaf community as it leads to the development of automatic recognition of emergencies from the sign language gestures. The sample video sequences from the dataset are shown in Fig. 5.

The gesture videos in this dataset were collected from 26 individuals (including 12 males and 14 females) in the age group of 22 to 26 years with two samples from each individual in an indoor environment with normal lighting conditions. The background of the videos was set as plain black color to avoid the presence of other moving objects from the scene. The original 520 video samples of the gestures were further replicated through image cropping-based data augmentation technique to get a total of 600 sample videos in the dataset with 60 samples for each category. The dataset is divided into two subsets with 50% data for training and the remaining 50% data for testing.

Classification has achieved an average accuracy of 97%. The overall performance of the classifier is expressed with the values of precision, recall, and f-score as given in Table 2. The plain black color background with uniform lightings in the videos of this dataset helps to increase the performance of HGR even with fewer training samples.

4 Experimental Study

The detailed descriptions of the experimental study conducted for the proposed ISL word recognition and the analysis of the results are described in this section.

4.1 Proposed Dataset of ISL Words Used by Deaf COVID-19 Patients

SLR aims to develop robust and efficient methodologies for the recognition of sign language gestures. The huge size of the sign language vocabulary and the regional variations that exist in the appearance of the gestures hinder the development of a universal SLR system. So the researchers focus on developing machine learning-based SLR models to be applied in specific domains. But one of the key challenges involved in this task is the lack of sufficient data samples for

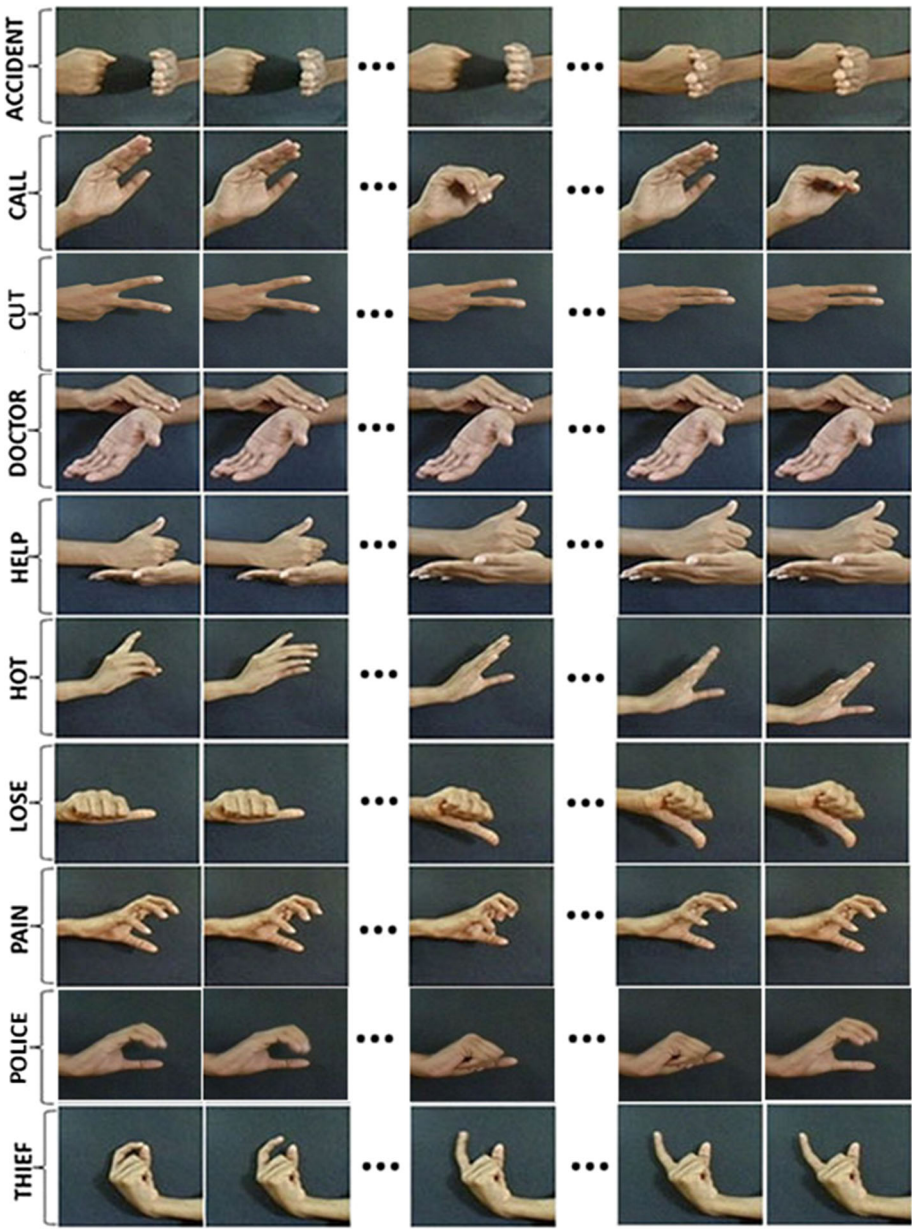

Fig. 5 The frame sequences of the hand gesture videos corresponding to the ISL words "accident," "call," "cut," "doctor," "help," "hot," "lose," "pain," "police," and "thief," respectively

Table 2 The classification performance of the proposed CNN-LSTM model on the ISL word dataset for emergency situations

ISL Word	Precision (%)	Recall (%)	F-score (%)
Accident	100	100	100
Call	93.33	93.33	93.33
Cut	100	100	100
Doctor	100	100	100
Help	100	100	100
Hot	100	96.67	98.31
Lose	93.55	96.67	95.08
Pain	100	86.67	92.86
Police	85.29	96.67	90.63
Thief	100	100	100

Fig. 6 The frame sequences of the hand gesture videos corresponding to the ISL words "breath," "call," "cough," "difficult," "distance," and "doctor," respectively

training and testing. The dataset creation by incorporating all the gestures in a particular domain is a very tedious and time-consuming process. Moreover, the people often show reluctance to pose in front of the camera as they are anxious about publishing their data. Many of the existing researches were carried out by building small datasets including limited numbers of participants.

The spread of the COVID-19 pandemic in India has increased the demand for developing an automatic SLR system that eases the communication of the COVID-19 positive deaf people staying in social isolation. With this objective, a video dataset of the hand gestures for the most common ISL words used to convey the COVID-19-related symptoms and emergencies has been created for the proposed work. The dataset includes videos of 17 hand gestures correspond to the ISL words "breath," "call," "cough," "distance," "difficult," "doctor," "help," "hungry," "lose," "pain," "smell," "taste," "temperature," "thirsty," "tired," "vomit," and "wash."

Five people including two men and three women from the age group of 25 to 55 years with different skin tones have participated in the data collection. Author's prepared an informed consent form (ICF) that clearly explains the objective of the research and the details of the data collection procedure. The ICF ensured the participants that their data will not be misused in any way. Moreover, they were given the awareness that the photographs and videos of only the "hands" will be captured to conduct this research, and the other parts of the body like the face that reveal their personality will not be published or exposed in any way. It is also clearly stated in ICF that their participation is voluntary and there is no foreseeable risk for them to participate in this study. All the participants have gone through the detailed ICF and signed their consent for voluntary participation. This data collection has got ethical clearance from the Institutional Human Ethics Committee (IHEC) of the Central University of Kerala, India.

The proposed dataset provides realistic gesture videos for developing SLR models. Two sample videos were collected from each participant in sitting as well as in standing positions getting a total of 340 ($17 \times 5 \times 2 \times 2$) original videos. The speed and style of hand movements vary for each participant resulting in gesture videos with unequal numbers of frame sequences. All the videos were captured from different indoor environments on different days at different time instances without imposing any restrictions on the backgrounds and illuminations. The sample frame sequences of the videos corresponding to each gesture category in the proposed dataset are shown in Figs. 6, 7, and 8, respectively.

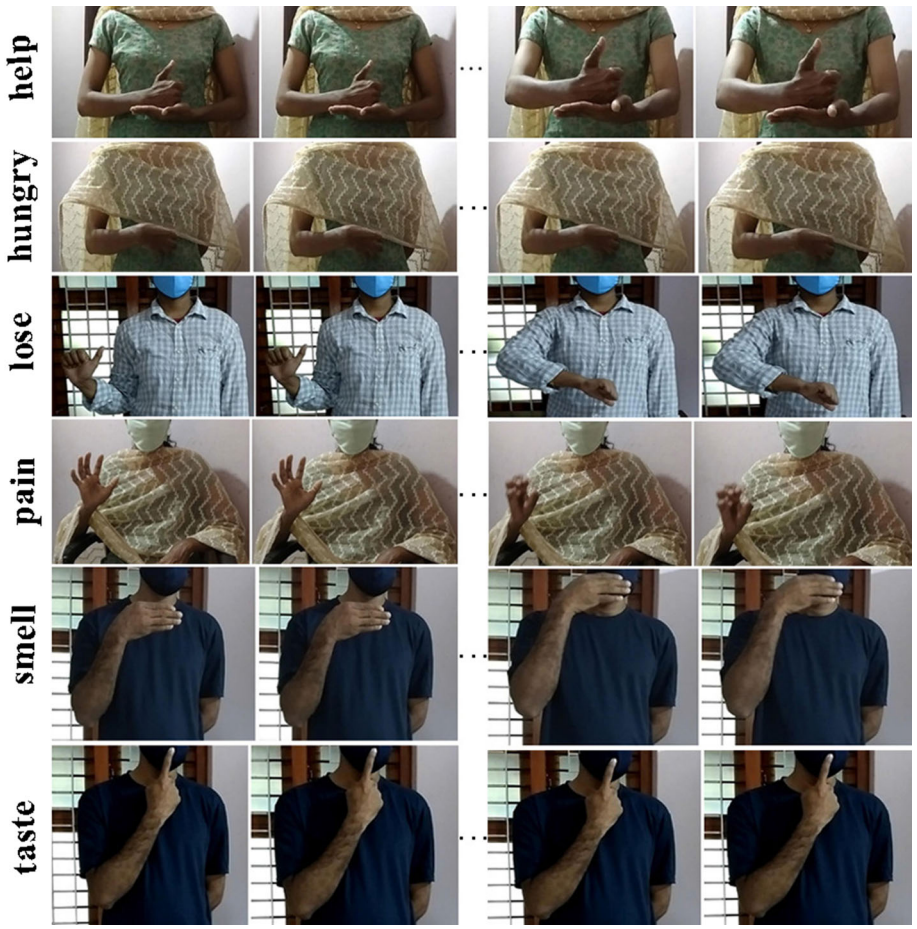

Fig. 7 The frame sequences of the hand gesture videos corresponding to the ISL words "help," "hungry," "lose," "pain," "smell," and "taste," respectively

Fig. 8 The frame sequences of the hand gesture videos corresponding to the ISL words "temperature," "thirsty," "tired," "vomit," and "wash," respectively

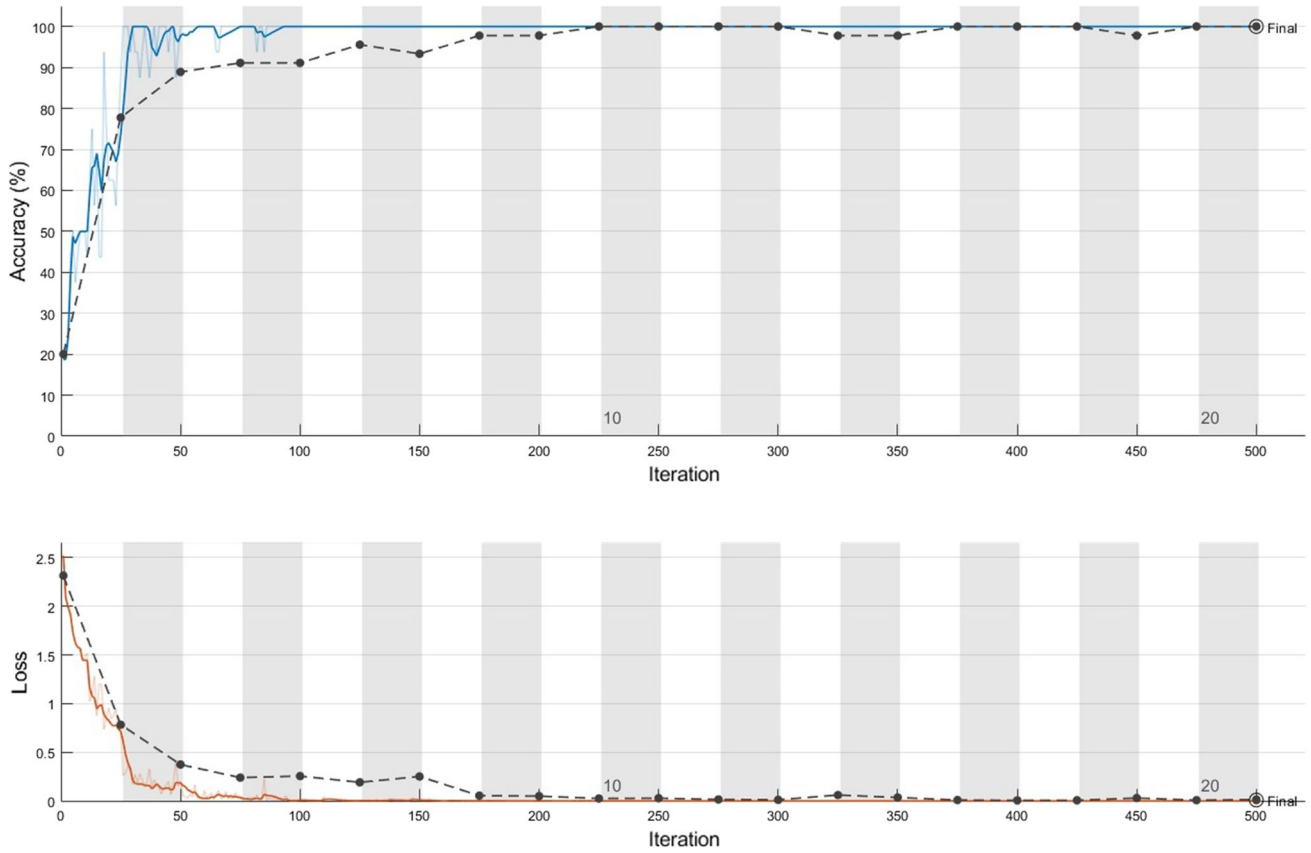

Fig. 9 A plot of the accuracy loss function for training the VGG16-BiLSTM network

4.2 Experimental Analysis of the Proposed ISL Word Recognition

The deep classification model defined with VGG-16 network and BiLSTM sequence network has been exploited for classifying the gestures in the proposed ISL dataset. The original videos of the hand gestures were replicated further with data augmentation through the image cropping technique. The individual frames of the videos were cropped on the left side, right side, and both left and right sides to make the hands appear in different positions in different proportions in the images. As the data collection has been carried out in real environments, the cropped samples show many differences in position and background scenes.

The data augmentation creates a dataset of 1360 video samples from the original 340 samples with 80 samples for each gesture class. The dataset is further divided into two halves with 50% data in the train set and the remaining 50% data in the test set. The frames of the gesture video samples are resized to ($224 \times 224 \times 3$) to match with the input size of the VGG-16 network. The resized frames are passed to the VGG-16 network consisting of stacked convolutional and pooling layers. Figure 2 shows the transformation of an input frame through the different layers of the VGG-16 network.

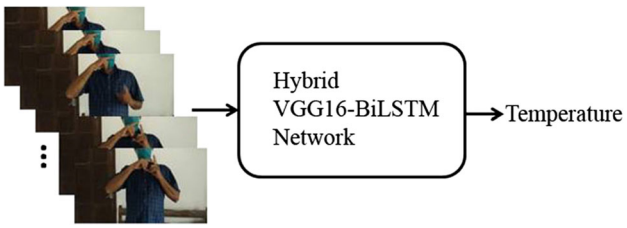

Fig. 10 An example for a gesture video input for the ISL word "temperature" and its corresponding text output

The output values from the last max-pooling layer are taken as the feature values of an image frame. Thus, the size of the feature vector is $25088 \times N$ for each video sample, where N is the number of frames.

In the training phase, the feature vector sequences obtained from the train set are passed to the BiLSTM classification network. The classifier is trained with 90% of the train data, in 20 epochs with the adaptive momentum (adam) optimization function, the initial learning rate of 0.0001, and the gradient threshold of value two. The remaining 10% of the train data has been utilized for validating the model. A plot of the accuracy loss function in training the VGG16-BiLSTM model is given in Fig. 9.

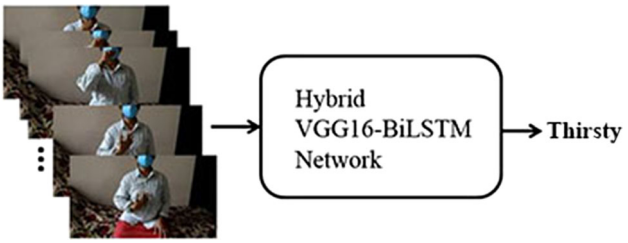

Fig. 11 An example for a gesture video input for the ISL word "thirsty" and its corresponding text output

In the testing phase, the sequences of feature vectors extracted from the test videos have been classified with the trained BiLSTM network. Figures 10 and 11 depict the application of the proposed classification model.

The classification has achieved an average classification accuracy of 83.36%. It means some of the gesture videos in the test set have been classified into wrong classes. The reason for misclassification may be the inconsistency of spatial and temporal features due to the similarity in the appearance of gestures. For example, among the 40 videos correspond to the gesture class "breath", 10 were classified into other gesture classes like "call," "pain," and "temperature" giving a failure rate of 25%. The hand gesture for "breath" is formed by moving the hand up and down near the nose and it is similar to other gestures like "call," "pain," and "temperature" that are also formed by some kinds of hand motions. In some cases, these gestures may resemble the motion patterns due to the highly flexible nature of human hands and the difference in hand movements shown by different individuals making the gesture features inconsistent.

The confusion matrix for classification is shown in Fig. 12. Classification performance is also evaluated in terms of statistical measures of precision, recall, and f-score values for each gesture category calculated as in Eqs. (7), (8), and (9), where TP is the true positive rate, FP is the false positive rate, and FN is the false negative rate, respectively.

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

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

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

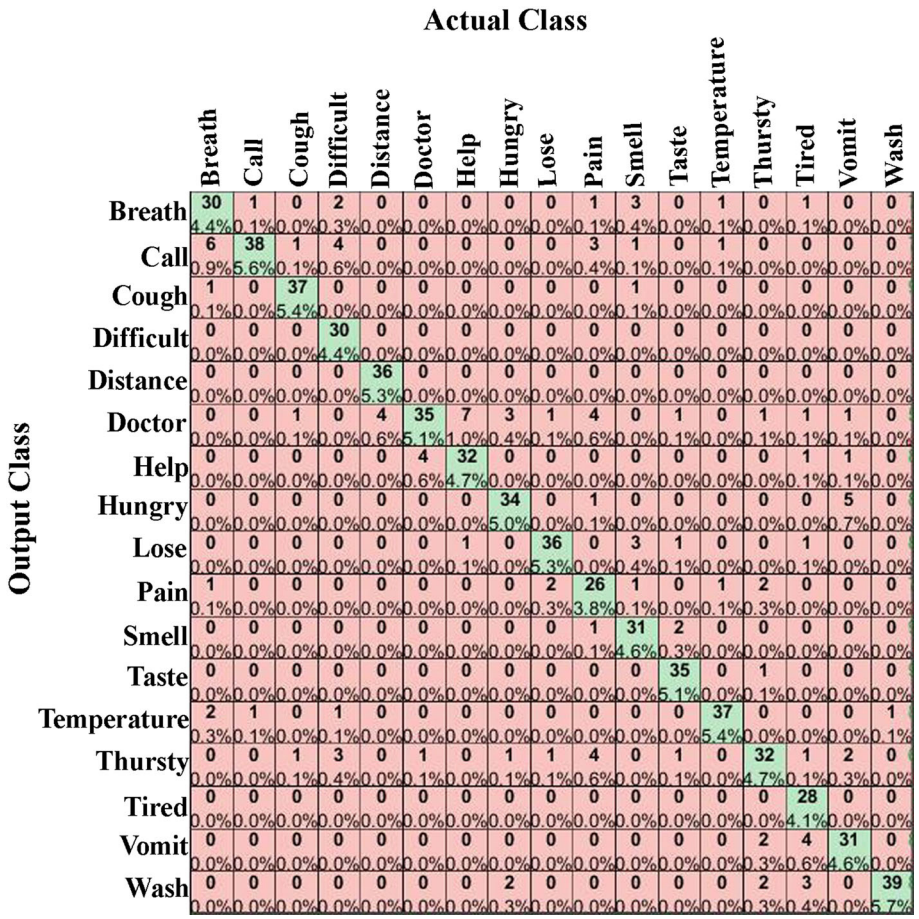

Fig. 12 Confusion matrix showing the classification performance of the proposed CNN-BiLSTM model on the ISL dataset of COVID-19-related words

Table 3 Classification performance of the proposed CNN-BiLSTM model on the ISL dataset of COVID-19-related words

ISL Word	Precision (%)	Recall (%)	F-score (%)
Breath	76.92	75	75.95
Call	70.37	95	80.85
Cough	94.87	92.5	93.67
Distance	100	75	85.71
Difficult	100	90	94.74
Doctor	59.32	87.5	70.71
Help	84.21	80	82.05
Hungry	85	85	85
Lose	85.71	90	87.8
Pain	78.79	65	71.23
Smell	91.18	77.5	83.78
Taste	97.22	87.5	92.11
Temperature	88.10	92.5	90.24
Thirsty	68.09	80	73.56
Tired	100	70	82.35
Vomit	83.78	77.5	80.52
Wash	84.78	97.5	90.70

The precision value indicates the measure of the positive identifications 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 overall performance of the classification model. The balanced precision and recall measures with the highest f-score measures reflect the performance of an optimal classification model. The overall classification performance of the hybrid CNN-BiLSTM model on the proposed ISL word dataset is shown in Table 3.

The experimental study has been conducted on a 3.3 GHz Intel® Xenon® W-2155 CPU with 32 GB memory. The proposed HGR model has been developed with raw gesture videos having unequal numbers of frames; the processing times of each step show considerable variations among the video sequences, i.e., the time taken for feature extraction is 1.5 ± 0.5 seconds and for classification is 1.5 ± 0.5. Possible solutions to further reduce the processing times for feature extraction and classification is to use GPUs (Graphical Processing Units).

Despite these promising results improvements are still needed in HGR as well as SLR research as it faces a lot of challenges like poor performance with increase in the number of signers belonging to different regions, ethnicities, handling of different skin colors, hand sizes and style of hand movements, handling more gesture classes with ambiguous motion patterns, recognition of sign language sentences etc. that open a wide opportunity for further research in this field. The techniques like rough set, fuzzy set, and Pythogorean fuzzy set [53] can be utilized to solve the challenges involved in classifying ambiguous and vague gesture movements and patterns. However, as an SLR model is trained with a set of specific sign language gestures, there may be chances that the users give wrong gestures that lead to the failure of the recognition system. Such problems can be solved by providing a menu display showing the correct style and structure of the recognizable gestures along with the SLR system.

5 Conclusion

The paper reports the automatic recognition of a set of dynamic hand gestures for the Indian sign language words commonly used by the deaf COVID-19 patients for emergency communication. Even though SLR has been addressed widely for normal times, this is the first work that focuses on the communication challenges of deaf people in the current pandemic situation. The videos of the hand gestures for the most common ISL words used for urgent communication by the COVID-19 positive deaf patients are included in the proposed dataset. The classification of the gestures is done with a hybrid model of VGG-16 and BiLSTM networks and achieved an average accuracy of 83.36%. The model has also been applied on another ISL word dataset as well as Cambridge hand gesture dataset to further assess its performance and achieved promising accuracies of 97% and $99.34 \pm 0.66\%$, respectively.

Still there are many unattended challenges in improving the accuracy of SLR when dealing with, gesture data belonging to a wide variety of gesture classes shown by signers from different regions of the world, recognition, and translation of continuous gesture sequences, occluded and ambiguous ges-

tures, and recognition of sign languages formed with hands, face, and body gestures. Hence, the future research directions indicate numerous potential opportunities in this field that will lead to the improvements in automatic SLR for a better living condition of the deaf community.

Acknowledgements 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 also like to acknowledge all the individuals who have participated in the data collection process.

References

1. Adithya, V.; Rajesh, R.: Hand Gestures for emergency situations: a video dataset based on words from Indian sign language. Data Brief (2020). https://doi.org/10.1016/j.dib.2020.106016
2. Wadhawan, A.; Kumar, P.: Sign language recognition systems: a decade systematic literature review. Arch. Comput. Methods Eng. (2017). https://doi.org/10.1007/s11831-019-09384-2
3. Elakkiya, R.: Machine learning based sign language recognition: a review and its research frontier. J. Ambient Intell. Humaniz. Comput. (2020). https://doi.org/10.1007/s12652-020-02396-y
4. Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. Artif. Intell. Rev. (2020). https://doi.org/10.1007/s10462-020-09825-6
5. Sherstinsky, A.: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys. D Nonlinear Phenom. (2020). https://doi.org/10.1016/j.physd.2019.132306
6. Pisharady, P.K.; Saerbeck, M.: Recent methods and databases in vision based hand gesture recognition: a review. Comput. Vis. Image Underst. **141**, 152–165 (2015). https://doi.org/10.1016/j.cviu.2015.08.004
7. Cheok, M.J.; Omar, Z.; Jaward, M.H.: A review of hand gesture and sign language recognition techniques. Int. J. Mach. Learn. Cybern. **10**, 131–153 (2019). https://doi.org/10.1007/s13042-017-0705-5
8. Matetelki, P.; Pataki, M.; Turbucz, S.; Kovacs, L.: An assistive interpreter tool using glove based hand gesture recognition. In: IEEE Canada International Humanitarian Technology Conference—(IHTC); Montreal, QC, Canada; (2014). pp.1-5. https://doi.org/10.1109/IHTC.2014.7147529
9. Soli, Google: Project soli, Google. https://atap.google.com/soli (2015)
10. Jaramillo-Yánez, A.; Benalcázar, M.E.; Mena-Maldonado, E.: Real-time hand gesture recognition using surface electromyography and machine learning: a systematic literature review. Sensors **20**, 2467 (2020). https://doi.org/10.3390/s20092467
11. Adaloglou, N.; Chatzis, T.; Papastratis, I.; Stergioulas, A.; Papadopoulos, G.T. et al.: A Comprehensive Study on Sign Language Recognition Methods. arXiv e-prints 2020; arXiv:2007.12530
12. Rastgoo, R.; Kiani, K.; Escalera, S.: Sign language recognition: a deep survey. Expert Syst. Appl. (2021). https://doi.org/10.1016/j.eswa.2020.113794
13. Jiang, X.; Zhu, Z.; Zhang, M.: Recognition of chinese finger sign language via gray-level co-occurrence matrix and k-nearest neighbor algorithm. In: 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE); Xiamen, China; 2019. pp. 152-156. https://doi.org/10.1109/EITCE47263.2019.9094915
14. Nogales, R.; Benalcazar, M.: Real-time hand gesture recognition using the leap motion controller and machine learning. In: IEEE Latin American conference on computational intelligence (LA-CCI); Guayaquil, Ecuador; (2019). pp.1-7. https://doi.org/10.1109/LA-CCI47412.2019.9037037
15. Verma, B.; Choudhary, A.: Grassmann manifold based dynamic hand gesture recognition using depth data. Multimed. Tools Appl. **79**, 2213–2237 (2020). https://doi.org/10.1007/s11042-019-08266-w
16. Neyra-Gutiérrez, A.; Shiguihara-Juárez, M P.: Feature extraction with video summarization of dynamic gestures for peruvian sign language recognition. In: IEEE XXVII International conference on electronics, electrical engineering and computing (INTERCON); Lima, Peru, (2020). pp.1-4. https://doi.org/10.1109/INTERCON50315.2020.9220243
17. Li, C.; Bai, X.; Xie, X.; Tian, L.: Dynamic hand gesture recognition based on 3D skeleton. In: IEEE 5th International conference on computer and communications (ICCC); Montreal, QC, Canada, (2019). pp. 1701-1705. https://doi.org/10.1109/ICCC47050.2019.9064200
18. Bai, X.; Li, C.; Tian, L.; Song, H.: Dynamic Hand gesture recognition based on depth information. In: International Conference on control, automation and information sciences (ICCAIS); Hangzhou, China, (2018). pp. 216-221. https://doi.org/10.1109/ICCAIS.2018.8570336
19. Agab, S.E.; Chelali, F.Z.: Dynamic Hand Gesture Recognition based on Textural Features. In: International Conference on Advanced Electrical Engineering (ICAEE); Algiers, Algeria; 2019. pp.1-6. https://doi.org/10.1109/ICAEE47123.2019.9014683
20. Ahmed, W.; Chanda, K.; Mitra, S.: Vision based Hand Gesture Recognition using Dynamic Time Warping for Indian Sign Language. In: IEEE international conference on information science (ICIS); Kochi, India, (2016). pp.120-125. https://doi.org/10.1109/INFOSCI.2016.7845312
21. Rao, G.A.; Kishore, P.V.V.: Selfie Video based continuous Indian sign language recognition system. Ain Shams Eng. J. **9**(4), 1929–1939 (2018). https://doi.org/10.1016/j.asej.2016.10.013
22. Athira, P.K.; Sruthi, C.J.; Lijiya, A.: A signer independent sign language recognition with co-articulation elimination from live videos: an Indian scenario. J. King Saud Univ. Comput. Inf. Sci. (2019). https://doi.org/10.1016/j.jksuci.2019.05.002
23. Ji, Y.; Zhang, H.; Jie, Z.; Ma, L.; Wu, Q.M.: Jonathan: CASNet: a cross-attention siamese network for video salient object detection. IEEE Trans. Neural Netw. Learn. Syst. **32**(6), 2676–2690 (2021). https://doi.org/10.1109/TNNLS.2020.3007534
24. Li, X.; He, M.; Li, H.; Shen, H.: A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection. IEEE Geosci. Remote Sens. Lett. (2021). https://doi.org/10.1109/LGRS.2021.3098774
25. Al-Hammadi, M.; Muhammad, G.; Abdu, W.; Alsulaiman, M.; Bencherif, M.A.; et al.: Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation. IEEE Access **8**, 192527–192542 (2020). https://doi.org/10.1109/ACCESS.2020.3032140
26. Aly, S.; Aly, W.: DeepArSLR: a novel signer-independent deep learning framework for isolated arabic sign language gestures recognition. IEEE Access **8**, 83199–83212 (2020). https://doi.org/10.1109/ACCESS.2020.2990699
27. Rahim, M.A.; Shin, J.; Islam, M.R.: Dynamic Hand Gesture Based Sign Word Recognition Using Convolutional Neural Network with Feature Fusion. In: IEEE 2nd international conference on knowledge innovation and invention (ICKII); Seoul, Korea; 2019. pp. 221-224. https://doi.org/10.1109/ICKII46306.2019.9042600
28. Lim, K.M.; Tan, A.W.C.; Lee, C.P.; Tan, S.C.: Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. Multimed. Tools Appl. **78**,

19917–19944 (2019). https://doi.org/10.1007/s11042-019-7263-7
29. Erhu, Z.; Xue, B.; Cao, F.; Duan, J.; Lin, G.; et al.: Fusion of 2D CNN and 3D DenseNet for dynamic gesture recognition. Electronics **8**, 1511 (2019). https://doi.org/10.3390/electronics8121511
30. Hoang, N.N.; Lee, G.-S.; Kim, S.-H.; Yang, H.-J.: A real-time multimodal hand gesture recognition via 3D convolutional neural network and key frame extraction. In: International conference on machine learning and machine intelligence; Ha Noi Viet Nam; (2018). pp. 32-37. https://doi.org/10.1145/3278312.3278314
31. Dhingra, N.; Kunz, A.: Res3ATN - Deep 3D Residual Attention Network for Hand Gesture Recognition in Videos. In: International Conference on 3D Visio; Québec City, QC, Canada; (2019). pp. 491–501. https://doi.org/10.1109/3DV.2019.00061
32. Nunez, J.C.; Cabido, R.; Pantrigo, J.J.; Montemayor, A.S.; Velez, J.F.: Convolutional neural networks and long short-term memory for skeleton based human activity and hand gesture recognition. Pattern Recognit. **76**, 80–94 (2018). https://doi.org/10.1016/j.patcog.2017.10.033
33. Hakim, N.K.; Shih, T.K.; Kasthuri Arachchi, S.P.; Aditya, W.; Chen, Y.C.; et al.: Dynamic hand gesture recognition using 3DCNN and LSTM with FSM context-aware model. Sensors **19**(24), 5429 (2019). https://doi.org/10.3390/s19245429
34. Kamruzzaman, M.M.: Arabic sign language recognition and generating arabic speech using convolutional neural network. Wirel. Commun. Mobile Comput. (2020). https://doi.org/10.1155/2020/3685614
35. Li, G.; Tang, H.; Sun, Y.; Kong, J.; Jiang, G.; Jiang, D.; Tao, B.; Xu, S.; Liu, H.: Hand gesture recognition based on convolution neural network. Cluster Comput. **22**, 2719–2729 (2019). https://doi.org/10.1007/s10586-017-1435-x
36. Tasmere, D.; Ahmed, B.: Hand gesture recognition for bangla sign language using deep convolution neural network. 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI) (2020); 1-5. https://doi.org/10.1109/STI50764.2020.9350484.
37. Avola, D.; Bernardi, M.; Cinque, L.; Foresti, G.L.; Massaroni, C.: Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. IEEE Trans. Multimed. **21**(1), 234–245 (2018). https://doi.org/10.1109/TMM.2018.2856094
38. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
39. Karen, S.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 2014
40. Van Houdt, G.; Mosquera, C.; Napoles, G.: A review on the long short-term memory model. Artif. Intell. Rev. **53**, 5929–5955 (2020). https://doi.org/10.1007/s10462-020-09838-1
41. Zhang, H.; Huang, B.; Tian, G.: Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. Pattern Recognit. Lett. **131**, 128–134 (2020). https://doi.org/10.1016/j.patrec.2019.12.013
42. Kim, T-K.; Wong, S-F.; Cipolla, R.: Tensor canonical correlation analysis for action classification. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN (2007)
43. Kurmanji, M.; Ghaderi, F.: Hand gesture recognition from RGB-D data using 2D and 3D convolutional neural networks: a comparative study. J. AI Data Min. **8**(2), 177–188 (2020). https://doi.org/10.22044/jadm.2019.7903.1929
44. John, V.; Boyali, A.; Mita, S.; Imanishi, M.; Sanma, N.: Dep learning-based fast hand gesture recognition using representative frames. In: International Conference on Digital Image Computing: Techniques and Applications (DICTA) (2016) https://doi.org/10.1109/DICTA.2016.7797030.
45. Lui, Y.M.: Human gesture recognition on product manifolds. J. Mach. Learn. Res. **13**(1), 3297–3321 (2012)
46. Sanin, A.; Sanderson, C.; Harandi, M.T.; and Lovell, B.C.: Spatio-Temporal Covariance Descriptors for Action and Gesture Recognition. In: IEEE Workshop on Applications of Computer Vision (WACV) (2013), 103-110
47. Baraldi, L.; Paci, F.; Serra, G.; Benini, L.; Cucchiara, R.: Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops 702–707 (2014). https://doi.org/10.1109/CVPRW.2014.107
48. Chandra, S.; Jawahar, C.V.: Partial Least Squares kernel for computing similarities between video sequences. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR) (2012); 13-516
49. Souza, L.S.; Gatto, B.B.; Xue, J.; Fukui, K.: Enhanced Grassmann discriminant analysis with randomized time warping for motion recognition. Pattern Recognit. (2020). https://doi.org/10.1016/j.patcog.2019.107028
50. Zhao, Z.; Elgammal, A.: Information theoretic key frame selection for action recognition. BMVC (2008)
51. Hoang, N.N.; Lee, G-S; Kim, S.-H.; Yang, H.-J.: effective hand gesture recognition by key frame selection and 3D neural network [Internet]. Korean Institute of Smart Media. Korean Institute of Smart Media 2020; 9: 23–29. https://doi.org/10.30693/SMJ.2020.9.1.23
52. Tang, H.; Liu, H.; Xiao, W.; Sebe, N.: Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion. Neurocomputing **331**, 424–433 (2019). https://doi.org/10.1016/j.neucom.2018.11.038
53. Wang, L.; Garg, H.: Algorithm for multiple attribute decision-making with interactive archimedean norm operations under pythagorean fuzzy uncertainty. Int. J. Comput. Intell. Syst. **14**(1), 503–527 (2020). https://doi.org/10.2991/ijcis.d.201215.002