Towards Indian Sign Language Sentence Recognition using INSIGNVID: Indian Sign Language Video Dataset

Kinjal Mistree[1], Devendra Thakor[2]
Computer Engineering Department
C. G. Patel Institute of Technology, Uka Tarsadia University, Bardoli, India

Brijesh Bhatt[3]
Computer Engineering Department
Dharmsinh Desai Institute of Technology, Dharmsinh Desai University, Bardoli, India

***Abstract*—Sign language, a language used by Deaf community, is a fully visual language with its own grammar. The Deaf people find it very difficult to express their feelings to the other people, since the other people lack the knowledge of the sign language used by the Deaf community. Due to the differences in vocabulary and grammar of the sign languages, complete adoption of methods used for other international sign languages is not possible for Indian Sign Language (ISL) recognition. It is difficult to handle continuous sign language sentence recognition and translation into text as no large video dataset for ISL sentences is available. INSIGNVID - the first Indian Sign Language video dataset has been proposed and with this dataset as input, a novel approach is presented that converts video of ISL sentence in appropriate English sentence using transfer learning. The proposed approach gives promising results on our dataset with MobilNetV2 as pretrained model.**

Keywords*—Indian sign language; sign language recognition; pretrained models; transfer learning; vision-based approaches

I. Introduction

Sign language is primary language for Deaf people to communicate with each other as well as with hearing people. One common perception exists which says that though deaf students use sign language as their main language, it is difficult for them to read texts fluently in their country's home language because their vision isn't hindered. But the fact is that signers learn their country's language as a secondary language, and so they cannot read and write texts fluently. For example, an English sentence *I don't understand* in sign language can be *I understand no*. If a person's primary language is sign language, all the other words are seen as noise which makes writing and comprehension in other language harder for him [1].

Therefore, communication in sign language becomes natural and easy way for Deaf. However, understanding sign language is difficult for hearing person, that is why human interpreters are required in emergency situations. They provide interpreting service that is easy-to-use but it has major limitations because India has only around 300 certified interpreters [2]. Because of the small number of human interpreters, interpretation service is not always available in urgent situations, training sessions and educational systems. Another service which became popular in last decade is online interpretation services, but this service is highly dependent on Internet connections from human interpreter and signer's end. Moreover, this service is also dependent on availability of human interpreters.

Given a recognized shortage of interpreters, an effective way of communication system should be developed that bridges the gap between signers and hearing people. The hearing person thus will easily understand the message that a Deaf want to convey. Such an approach should be the best alternative to the human interpreter in case of emergency situations or even to access the services in local community. The system should be designed in such a way that it should also improve quality of life for Deaf community.

The main dimensions of research in automatic sign language recognition are grouped as: Recognition of Isolated signs and recognition of continuous sign. In isolated sign recognition, static or dynamic single signs are recognized without continuation of any other sign. Isolated sign is not affected by the previous or following sign. In continuous sign language recognition system, different signs are performed one after another to recognize complete sign language word or sentence. Considering sign languages at international level, many differences exist among them. The same datasets and methods cannot be used for Indian Sign Language (ISL) recognition due to the very few meaningful similarities in grammar and vocabulary of all the sign languages [3-5]. In India, major research is focused on regional versions or manual components of signs. [6-11] reported work on dynamic hand gesture recognition for their dataset having limited number of gestures for single and two-hands manual signs. For Indian sign language translation, [12] and [13] proposed algorithms to convert ISL sentences in English text but they used traditional rule-based approach for sign translation due to limited size of their datasets. Also, the existing datasets are not prepared according to dictionary released by Indian Sign Language Research and Training Centre (ISLRTC).

Compared to other international sign languages, Indian sign language recognition and translation is still in inception. Device or sensor-based and vision-based approaches have been used for recognizing sign language from images of sign. Vision-based approach is better than device-based methods as device-based methods need extended setup and they also limit the inherent movement of the face and hands.

If we see sign language recognition as image recognition problem, deep neural networks perform exceptionally well for such tasks [14]. But these networks are largely dependent on huge data to avoid problem of overfitting. Overfitting refers to the phenomenon when a network learns a function with very high variance such as to perfectly model the training data [2]. One of the ways for dealing with the problem of limited dataset size is called data augmentation [15]. This approach can be used on input videos for Indian sign language recognition to make dataset inflated.

In this article, one research question has been addressed in particular: how to use deep neural network with very small amount of input videos in order to incorporate both left-handed and right-handed signs without hurting recognition performance of ISL sentences. A dataset – INSIGNVID is also created, which is the first ISL video dataset that uses official ISL signs released by ISLRTC. On this dataset all steps of proposed approach are performed and promising results are shown using MobileNetV2 pretrained model. This article is organized as follows: Brief literature review of existing ISL recognition systems is discussed in Section 2. Detailed description of dataset-INSIGNVID is given in Section 3. This dataset is created to motivate research in field of ISL dynamic gesture recognition. Section 4 discusses steps of the proposed approach with detailed explanation. Section 5 describes implementation scenario and results of ISL gesture recognition using proposed approach. Section 6 provides concluding remarks and directions for future work.

II. Related Work

Considering sign language recognition, two types of different approaches exist: (i) Sensor or glove-based approach and (ii) Vision based approach. Sensor based methods have advantage of extracting the signer's movements and postures more accurately because they use specialized gloves, which are embedded with several sensors to capture the sign information. There is a significant amount of work reported in ISL recognition giving good accuracy with different methods using sensors, and a thorough review is presented in [16]. However, it is practically impossible to wear the gloves by signers in their daily activities as they restrict the movement of signers. Also, in the emergency situations, this setup may not be available with signer. Vision based approach on the other hand, is cost effective and flexible solution while touch based approach is complex, costly and difficult to deploy [17].

By taking different international sign language recognition systems into account, many standard datasets are available publicly. The author in [18] proposed RWTH-PHOENIX-Weather video dataset, for Germen sign language recognition, translation and pattern recognition. The same dataset was extended by [19], which has tripled the original dataset in size. Both of these datasets contain sign languages of news related to weather forecasting. SIGNUM dataset [20] was used for German sign language recognition to handle multiple signers through statistical approach. [21] formulated Bayesian network to recognize Boston American sign language using videos of American sign language lexicon video dataset (ASLLVD). Polish sign language word dataset – PSL Kinect 30 was created with Microsoft Kinect and was used by [22] with skin color-based features. The authors also performed the same experiments on another Polish sign language dataset – PSL ToF 84 and discussed the results on both of these datasets. [23] have proposed a pre-processed version of an Argentinian sign language dataset – LSA64 to promote research in Argentinian sign languages. Human action recognition using depth sequences of MSR Gesture 3D – American sign language was implemented by [24] using gradient features.

A significant amount of research is done on static sign recognition for ISL. The authors in [25-30] have proposed approaches for isolated ISL recognition. Microsoft Kinect became popular choice for work in sign language recognition when Kinect was launched in year 2010. Kinect traces motion in full body and has depth sensors that makes pose estimation of signer perfect. The author in [31] used Microsoft Kinect to recognize static as well as dynamic ISL signs using Multi-class Support Vector Machine (SVM). The features were extracted from 20 skeleton joints of a human body. The authors have achieved 86.16% accuracy on the test data. The author in [32] have proposed an approach to recognized single and double handed ISL signs using combination of Kinect and leap motion. Authors achieved 95.60% accuracy using Hidden Markov Model and Bi-directional Long-Short Term Memory (BLSTM) on 7500 gestures of ISL signs. Authors then used Coupled Hidden Markov Model (CHMM) with the concept of multi-fusion and achieved 90.80% accuracy in [33] for single handed signs. 83.77% accuracy was achieved by [34] for 2700 gestures with Kinect and leap motion. The authors used Coupled HMM to recognize single handed dynamic signs. Table I shows comparative analysis of vision-based Indian Sign Language gesture recognition systems considering camera as acquisition instrument.

The author in [8] proposed an approach to recognize ISL alphabets using double handed signs. Authors have created a dataset for 3 dynamic signs and created 60 videos. Authors have used YCbCr color segmentation and used Principle Curvature Based Region (PCBR) detector and Wavelet Packet Decomposition (WPD-2) methods with Dynamic Time Warping (DTW) classifier. Authors achieved 86.3% accuracy using Support Vector Machine (SVM).

TABLE I. REVIEW OF VISION-BASED DYNAMIC INDIAN SIGN LANGUAGE (ISL) RECOGNITION SYSTEMS

Year and Authors	Segmentation Technique	Feature Extraction Technique	Classification Technique	No. of Samples	Sign Types
Jayaprakash and Majumder, 2011	YCbCr	Principle Curvature Based Region + Wavelet Packet Decomposition-2 + convexity detect	Dynamic Time Warping	60	Alphabets
(Sahoo et al., 2019)	Skin color segmentation	Direct Pixel Value (DPV), local histogram, and hierarchical centroid (HC)	Artificial Neural Network	5000 (digits), 2340 (alphabets), 1250 (words)	Digits, alphabets and words
(Tripathi and Nandi, 2015)	HSV	Orientation histogram + Principle Component Analysis	Distance based classifier	60	Sentences
(Baranwal et al., 2017)	HSV, Otsu thresholding	Wavelet descriptor and Mel Sec Frequency Cepstral Coefficients (MFCC)	kNN, Support Vector Machine (SVM)	8	Words
(Kishore et al., 2016)	Horn Schunck Optical Flow	Active Contour model	Artificial Neural Network	1	Sentence having 58 words
Kishore and Anil (2016)	Sobel with adaptive block thresholding	Contour based model	Artificial Neural Network	18	Words
(Baranwal et al, 2017)	Background modelling	Wavelet descriptor	Possibility theory	20	Sentences
(Wazalwar and Shrawankar, 2017)	HSV	Pseudo 2-dimensional Hidden Markov Model	Harr classifier	60	Sentences
(Mariappan and Gomathi, 2019)	Skin segmentation feature of OpenCV	Regions of Interest	Fuzzy c-means clustering	130	Words, sentences
(Bhavsar et al., 2020)	Viola-Jones algorithm	Distance count and Correlation-Coefficient methods	Neuro-Fuzzy classifier	100	Words

The author in [35] created 1250 video for word signs, where each category has 125 videos. They have shown comparison of proposed approach with kNN and Artificial Neural Network (ANN) as classifier. For each video 2-5 frames are considered in sequence to identify words. The overall accuracy achieved using kNN classifier with DPV technique is 96.70% and with HC technique is 93.70%. This approach works for specific signer as ANN is trained on hand and face features. The author in [13] proposed continuous Indian sign language sentence recognition using various distance-based classifiers from which Euclidian and Correlation distance classifiers give accuracy up to 93%. Authors converted RGB frames in HSV frames using thresholding, from which hand portion was extracted. Meaningful sequence of frames was extracted by using gradient method. Hand features were extracted by using orientation histogram and then distance based classifiers were applied for classification. Segmentation of hand from upper half of the body and boundary changes depending on hand shape of various signers are solved by [11]. Authors have used OTSU thresholding method for segmentation, Mel Sec Frequency Cepstral Coefficients (MFCC) method for extracting features. But the algorithm doesn't work on all geometric and photometric transformation techniques. Backpropagation algorithm was used by [36] on continuous sign language sentences using active contour hand shape features. Though the authors achieved 90.172% by capturing videos from four different signers, linguistic rules are not considered while performing sign language gestures.

The concept of possibility theory was used by [10] on 20 different continuous gesture videos of single-handed signs. These videos are captured with various backgrounds and multiple objects were used with black cloths having full sleeves. Authors achieved 92% accuracy on their own dataset.

The author in [12] proposed an algorithm for converting ISL sentences to English text. Authors used Pseudo 2-dimensional Hidden Markov Model (P2DHMM) for feature extraction, which is proven better than simple Hidden Markov Model. For converting recognized signs in English text, LALR parser was used. Major limitation of this work is that signs for words are recorded and then they are combined to create sentences. Furthermore, the algorithm worked for signs performed by single hand only.

The author in [37] achieved 90% accuracy on continuous signs captured using front camera of mobile phone. 10 signers have performed the sign of words and these words are arranged in specific order to create sentences. At the time of testing, video with same words with different order are given as input. The author in [38] discussed approach based on Fuzzy c-means clustering for recognizing sign words using 800 samples from different 10 signers. The author in [39] proposed an approach to classify word signs using Neuro-Fuzzy approach. Authors displayed the final word using Natural Language Processing (NLP) technique and achieved 95% accuracy.

To sum up, sign language recognition is a challenging problem as the recognition task involves visual and lingual information interpretation. Most of the research that has been conducted in Indian sign language recognition till date has considered this task as a gesture recognition problem, and ignores the linguistic properties of the sign language and has assumed that there is word-to-word mapping of sign to speech. Less amount of work is reported considering dynamic gestures of ISL words and these approaches works on limited vocabulary of signs. Also, all the work described here has considered manual components of signs, ignoring facial expressions. Moreover, most of the video capturing is done in controlled laboratory settings. No video sentence dataset is publicly available that is developed using ISL signs released by ISLRTC. Looking into these limitations, we were motivated to work in the direction of creating new dataset and an algorithm that works in fully functional way on this dataset without hampering video recognition accuracy.

III. INSIGNVID: INDIAN SIGN LANGUAGE VIDEO DATASET

Currently no video dataset is publicly available for ISL sentence recognition using dictionary launched by ISLRTC, as discussed in Section 1. This section introduces dataset - INSIGNVID, containing set of video sequences of ISL sentences using word signs released by ISLRTC. Through recording it was observed that same signer can perform sign in slightly different manner each time; there may be variations in speed, gesture position and facial expressions. To incorporate these variations, signers have performed signs multiple times for each ISL sentence.

A. *Video Characteristics*

To create INSIGNVID, videos are captured from frontal view, using CANON EOS 700D digital SLR camera with 18 MP resolution. Videos are captured at 30 FPS (Frames Per Second), with a resolution of 1920 x 1088 pixels per frame. High resolution was used for recording videos because the facial expressions and hand movements would be captured in more detail in high resolution. This will facilitate researchers to convert the videos in low resolution, if required for specific applications.

Green background is used while recording the video because background removal is quite extensive task and it also affects the object recognition quality in real scenarios, if typical image processing techniques are used. If one wants to replace the background, a commonly used technique called as color keying can be used. Color keying can be done with any color having the background that is not matched with human skin color. Moreover, color chosen for color keying should be discrete and uniform. Green is commonly used color for color keying and it is generally preferred color over other colors as it requires less energy to produce uniform illumination over the color screen [40]. Due to this reason, all videos are captured using green backdrop. During the dataset creation, the distance between the camera and the signer is adjusted so that the dynamic gestures can be captured from upper part of the body. Later, using augmentation techniques, zooming and rotation can be performed to make generalized videos. While recording the videos of ISL sentences, proper lighting conditions are considered.

B. *Dataset Description*

55 most frequently used sentences were identified, after having discussion with teachers at Muk Badhir Seva Trust, Surat, India. All these sentences are composed of more than one word or hand strokes and are combination of single and double handed hand gestures. A total of 1289 videos of ISL sentences corresponding to 55 unique class labels have been captured with the help of signers. While creating dataset, signs were performed by a total of 4 right-handed signers. Green backdrop was used and black clothes were worn by signers while performing video. Signers of different age groups and both the genders have performed signs to make the system useful to all. To capture multiples variances in signs by each signer, the signs were captured 4-6 times from each signer. For capturing ISL videos of Deaf and authentication of signs performed, we have taken help and consent of ISL interpreter Ms. Mital Joshi, who is one of the certified ISL interpreters by ISLRTC. Table II shows some examples of English sentences and corresponding ISL sentences used for recording video sequences.

ISL is considered to be Subject-Object-Verb (SOV) language unlike English language which considers Subject-Verb-Object (SVO) order.

Sign languages do not use articles, linking verbs, prepositions of time like *for, by, during* etc. Fig. 1 shows sample frames of English sentence *Please wait for some time*. It can be observed from the frames that words please and wait are represented by single sign, while two signs are required to represent word time.

As shown in Fig. 2, sentence Look at the beautiful butterflies is represented as ISL sentence Beautiful butterflies look. Fig. 3 shows five sample frames of ISL sentence He look like his father for corresponding English sentence He looks like his father.

TABLE II. EXAMPLES OF ISL SENTENCES AND CORRESPONDING ENGLISH SENTENCES FROM DATASET-INSIGNVID

Sign Language Sequence	**English Sentences**
Her smile beautiful	Her smile is beautiful.
My daughter always happy	My daughter is always happy.
We late	We are late.
My brother restaurant work	My brother works in the restaurant.
Hello you how	Hello, how are you?
Wrong what	What's wrong?
You who	Who are you?
Tomorrow people come how	How many people will come tomorrow?
Thermometer where	Where's the thermometer?
I understand not	I don't understand
Problem no	No problem.
He banana eat no	He doesn't eat banana.
Good morning	Good morning
Thank you	Thank you
Hurry up we late	Hurry up, we are late!

Fig. 1. Sample ISL Frames for English Sentence Please Wait for Some Time.

Fig. 2. Sample ISL Frames for English Sentence Look at the Beautiful Butterflies.

Fig. 3. Sample ISL Frames for English Sentence he Looks Like his Father.

Fig. 4. Sample ISL Frames for English Sentence I Don't Understand.

Fig. 5. Sample ISL Frames for English Sentence What is Your Name?.

In case of negation clause, word representing negation is signed at the end of sentence. Fig. 4 shows signs for ISL sentence I understand no corresponding to English sentence I don't understand.

If the sentence is interrogative sentence, word representing question comes at the end of ISL sentence. For example, English sentence What is your name? is converted as Your name what in ISL, as shown in Fig. 5.

To sum up, the first ISL video dataset-INSIGNVID has been created, using ISL dictionary launched by ISLRTC. We hope that our video dataset will become a helpful resource to motivate research in field of gesture recognition, sign language recognition and human-computer interaction.

IV. Our Approach

The process of formation of English sentences from continuous ISL gestures mainly involves ISL video recognition. The design of the proposed approach is shown in Fig. 6 and the steps are discussed in detail below.

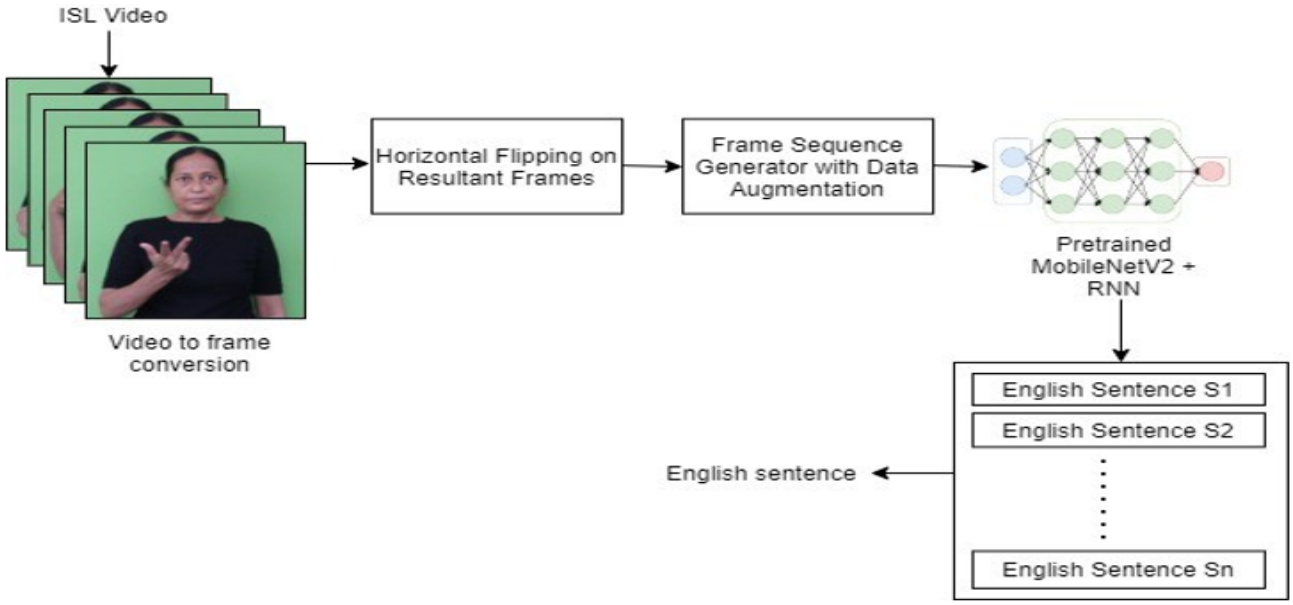


Fig. 6. Framework of Proposed Approach.

1) Video to frame conversion: Videos in dataset were captured using 30 fps, with resolution of 1920 x 1088 pixels. Taking these characteristics, RGB frames are generated for each video of different length.

2) Horizontal flipping: Signer can be either right-handed or left-handed. On the batch of frames, horizontal flipping is performed in order to incorporate both left-handed and right-handed signs. Though horizontal flipping is one of the data augmentation techniques, this technique has been explicitly separated from the other set. Reason behind this is, if online data augmentation technique is used, the modifications in images will be applied randomly, so not every image will be changed every time.

3) Frame sequence generator with image augmentation: To identify main frames, one possibility is that N distributed frames can be picked from the entire video. But this works only with fixed length of video as we may lose important information from frames. To address this issue, video generator is created that provides decomposed frames for the entire video in sequential form to get more sequences from one video file. For each 30 FPS video this video generator is used to select 5 frames per second. It has been decided to select every 6th frame based on analysis of histogram difference in frames. For each individual video, frames are selected in batches in order to get a set of shifted frames, such as first batch has frames 1, 7, 13, 19, 25 in sequence; second batch has frames 2, 8, 14, 20, 26 in sequence and so on. This custom generator supports image augmentation techniques. On the resultant images after frame sequence generator, geometric transformations - zooming, rotation, vertical shifting, horizontal shifting; and photometric transformations, augmentation on brightness are performed.

The author in [41] has discussed how to produce promising ways to increase the accuracy of classification tasks using data augmentation. It has been decided to work with augmentation techniques based on two aspects: Various video recording conditions and hardware dependency. For end-to-end ISL recognition, the environment in which signers perform signs under lighting and camera settings may be different. Signers may use different hardware devices such as camera, smartphone, tablets, computer with different resolutions and view. These variances are addressed by training the deep learning model with randomly selected augmentation types within range of parameters. It has been shown that training the recognizer with inflated data with randomness in augmentation gives remarkable improvement in accuracy. Image augmentation types and parameters were randomly selected with frame sequence generator.

4) Training with MobileNetV2 + RNN: Video frame generator is created that acts as video generator. We have taken 5 frames per second for each video. The inflated dataset after image augmentation technique is given as input to the Convolutional Neural Network (CNN). CNN architecture is selected for our work as it is best suited for capturing internal representation of features of visual world [42]. Here, our CNN is initiated with MobileNetV2 [43] model that was pretrained on ImageNet dataset. Pretrained model is chosen because image augmentation increases the size of the dataset which is originally very small but the data similarity is still very high. Also, MobileNetV2 is light-weight model that uses deep neural network that has proven best for mobile and embedded vision applications [43]. Fine-tuning is performed on the MobileNetV2 model by experimentally changing the top layer configuration of the model to get the best classification result. Moreover, LSTM (Long-Short Term Memory) model has been added that needs one dimension. As MobileNetV2 model is used without top layers, one Time Distributed layer has been added as a top layer to have the one-dimension shape compatible with LSTM. Finally, dense layer is added to get the prediction of English sentence. Output of overall process will be semantically equivalent English sentence corresponding to ISL sentence.

V. Experiments on INSIGNVID: Indian Sign Video Dataset

A set of experiments is conducted on our dataset - INSIGNVID, using our proposed approach. As discussed in Section 3.2, a total of 1289 videos have been used for training, validation and testing our model. Total videos are divided in three parts: 70% videos are used for training, 15% videos for validation and 15% videos for testing. The data used for testing has signs of the fourth signer which were not used for training and validation of the model. Therefore, the total number of training, validation and testing samples after horizontal flipping were 1808, 384 and 386, respectively. Frame sequence generator generates 5 video frames per second on which image augmentation techniques are performed. As a result, a total of 52432, 16848 and 16796 sequences are generated for training, validation and testing, respectively. Table III shows the statistics of image sequences after using frame sequence generator with data augmentation techniques.

The performance of MobileNetV2 model is also compared with three popular pretrained models used for object recognition: ResNet50, VGG16 and MobileNet. Table IV shows the comparison of model performance of MobileNet, MobileNetV2, ResNet50 and VGG16 models on.

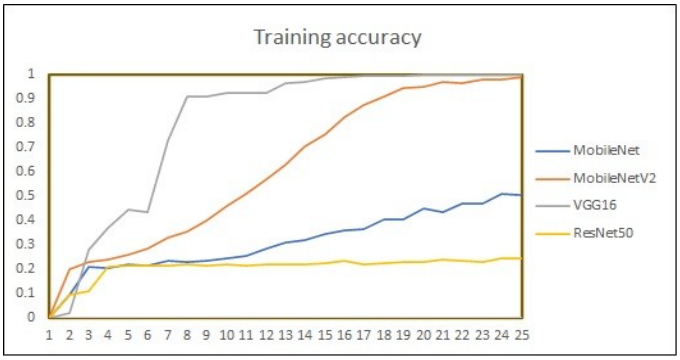
INSIGNVID after 25 epochs: Plots of accuracy and loss of all pretrained models are shown in Fig. 7. Here, it is important to note that even after 40 epochs (nearly after 35 hours), ResNet50 could not achieve accuracy more than 57%. VGG16 model achieves highest training accuracy but could get 89% testing accuracy. MobileNet model achieved 91% accuracy after 21 hours because it is slower than MobileNetV2 model.

TABLE III. IMAGE SEQUENCES AFTER USING FRAME SEQUENCE GENERATOR WITH DATA AUGMENTATION ON INSIGNVID

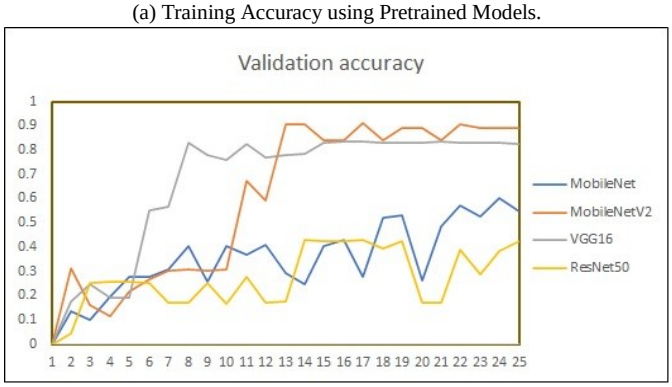
Parameters	**Values**
No. of classes (ISL words)	55
Average video length (seconds)	5
Total samples	1289
Training samples	904
Validation samples	192
Testing samples	193
Training samples after horizontal flipping	1808
Validation samples after horizontal flipping	384
Testing samples after horizontal flipping	386
No. of training sequences after using frame sequence generator	52432
No. of validation sequences after using frame sequence generator	16848
No. of testing sequences after using frame sequence generator	16796

TABLE IV. COMPARISON OF PRETRAINED MODELS' PERFORMANCE ON INSIGNVID

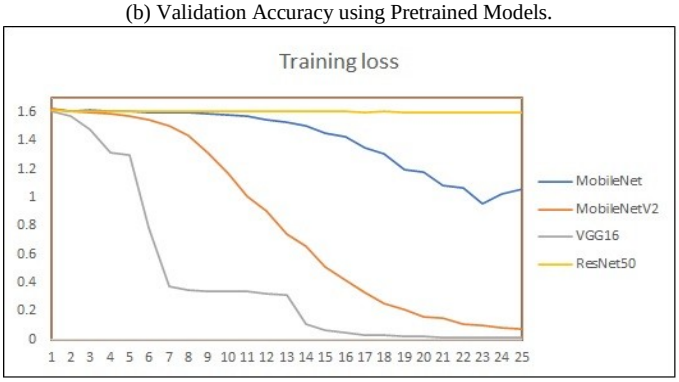
Parameters	**MobileNet**	**ResNet50**	**VGG16**	**MobileNetV2**
Model size (MB)	16	98	528	14
Trainable layers	9	9	9	9
Total Parameters	4,173,253	24,794,245	14,714,688	3,267,909
Non-trainable parameters	2,166,976	22,531,968	5,275,456	1,537,984
Trainable parameters	2,006,277	2,262,277	9,439,232	1,729,925
Time in hours (training + validation)	19.76	23.04	17.79	14.6
Training accuracy (%)	50.1	23.06	99.92	99.04
Validation Accuracy (%)	52.35	41.97	80.06	90.31
Testing accuracy (%)	50.4	32.09	89.06	93.89
Training loss	1.1068	1.60	0.0547	0.1981
Validation loss	1.1869	1.5875	0.953	0.2092
Testing loss	1.087	1.2624	0.1521	0.1735
Time in hours to get >=80% accuracy	23	Max. 57% after 35 hours	12.7	12.1



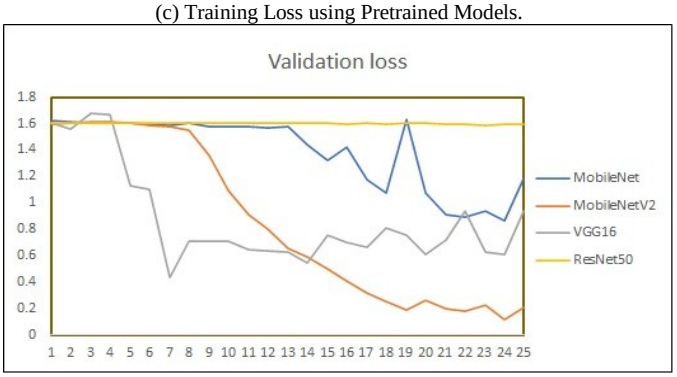
(a) Training Accuracy using Pretrained Models.



(b) Validation Accuracy using Pretrained Models.



(c) Training Loss using Pretrained Models.



(d) Validation Loss using Pretrained Models.

Fig. 7. Plots of Accuracy and Loss using Pretrained Models on INSIGNVID.

By looking at the comparison, it is clear that MobileNetV2 outperforms all other models for our dataset. 93.89% testing accuracy was achieved after 726 minutes, using MobileNetV2 model. Table V shows the hyperparameters used for all described models. Fig. 8 and Fig. 9 summarizes overall performance of proposed approach on our dataset - INSIGNVID. For calculation of precision, recall and F1-score, the following formulas are used [44]:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (1)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (2)$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

TABLE V. HYPERPARAMETERS CONFIGURATION FOR PERFORMANCE COMPARISON BETWEEN PRETRAINED MODELS

Model configuration	Parameters
Model	Pretrained model + LSTM
Trainable layers	9
Learning rate	0.002
Dropout	0.4
Batch size	128
Activation function	ReLU
Optimization algorithm	Stochastic gradient descent

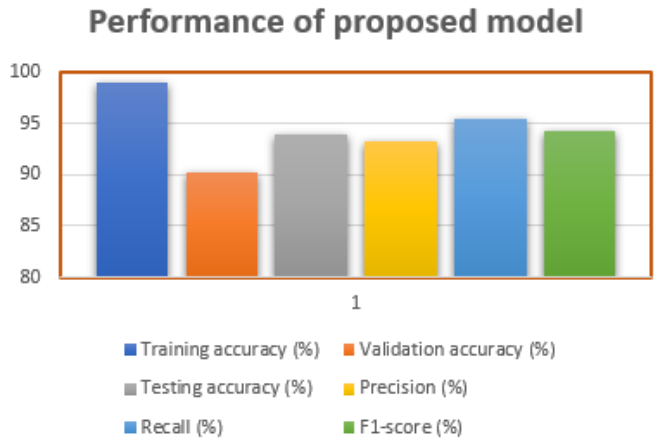
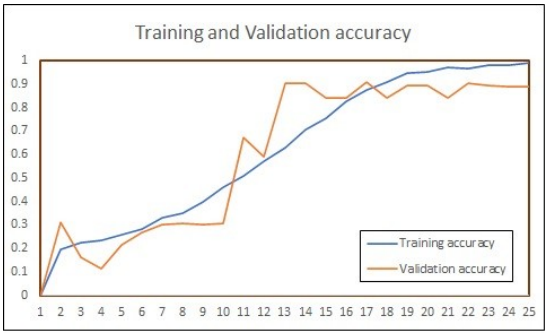
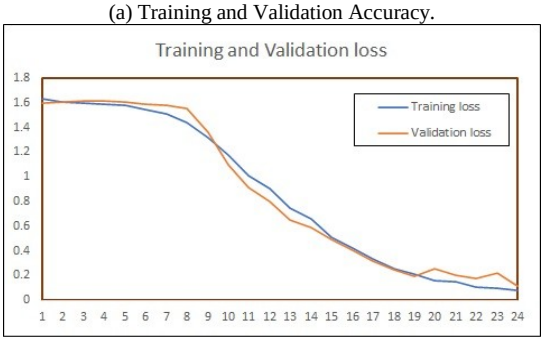


Fig. 8. Performance of Proposed Approach on INSIGNVID.



(a) Training and Validation Accuracy.



(b) Training and Validation Loss.

Fig. 9. Plots of Accuracy and Loss using MobileNetV2 on INSIGNVID.

VI. Conclusion and Future Work

Developing systems that translates sign language video in sentences is still a challenging task in India, as no video dataset for ISL sentences are available publicly. The existing datasets depend on local version of signs, have limited categories of gestures and have high variations in sign. At the same time, it is essential to work on continuous ISL sentence recognition, as it is primary language used by Deaf community to communicate with other people. Also, it becomes a challenging task when one wants to achieve more accuracy with a smaller number of samples in generalized environment. Deep learning gives promising results than other traditional algorithms in computer vision task as they learn features from gestures, but they require huge dataset. To overcome the problem of overfitting generated by deep learning models on less amount of data, image augmentation can be used before training data which also increases accuracy of test data. In this work, it has been empirically proven that simple image manipulation techniques and pretrained model with frame sequence generator creates great impact on the accuracy on ISL recognition than using very limited amount of data in training. An approach is proposed that uses pretrained model MobileNetV2 on our new dataset – INSIGNVID. This model learns features from augmented frame sequences of ISL gestures using batch of shifted frames to provide decayed sequences for the same gesture. Using MobileNetV2 as pretrained model on our dataset, promising results are shown for ISL sentence recognition.

In future, the proposed approach will be tested against unseen sentences. Furthermore, machine translation approach will be studied and implemented on parallel corpora of English and ISL sentences. The ISL corpus will be used for testing ISL sentences and the performance will be evaluated with evaluation parameters.

References

[1] Oliveira, Tiago & Escudeiro, Paula & Escudeiro, Nuno & Rocha, Emanuel & Maciel-Barbosa, Fernando. (2019) 'Automatic Sign Language Translation to Improve Communication' in *2019 IEEE Global Engineering Education Conference (EDUCON)*. pp.937–942.

[2] Ghotkar, Archana & Kharate, Gajanan. (2015) 'Dynamic Hand Gesture Recognition for Sign Words and Novel Sentence Interpretation Algorithm for Indian Sign Language Using Microsoft Kinect Sensor', Journal of pattern recognition research, Vol. 1, pp.24–38.

[3] Verma, Vivek & Srivastava, Sumit. (2018) 'Towards Machine Translation Linguistic Issues of Indian Sign Language', Speech and Language Processing for Human-Machine Communications, Vol. 664, pp.129-135.

[4] Zeshan U. (2013) 'Distinctive features of Indian Sign Language in comparison to foreign sign languages', The People's Linguistic Survey of India.

[5] Zeshan, U. (2006) 'Sign languages of the world' in Encyclopedia of Languages and Linguistics, Amsterdam, The Netherlands: Elsevier, pp. 358-365. http://clok.uclan.ac.uk/9631/.

[6] BM, Chethankumar & Chinmayi R, Lekha. (2016) 'Indian Sign Language Recognition: An Approach Based on Fuzzy-Symbolic Data' in International Conference on Advances in Computing, Communications and Informatic, pp.1006-1013.

[7] Gupta, Bhumika & Shukla, Pushkar & Mittal, Ankush. (2016) 'K-nearest correlated neighbor classification for Indian sign language gesture recognition using feature fusion' in International Conference on Computer Communication and Informatics (ICCCI), pp.1–5.

[8] Jayaprakash, Rekha & Majumder, Somajyoti. (2011) 'Shape, Texture and Local Movement Hand Gesture Features for Indian Sign Language Recognition' in 3rd International Conference on Trendz in Information Sciences & Computing (TISC2011), pp.30-35.

[9] Sahoo, Ashok & Mishra, Gouri & Ahmed, Pakhshan. (2012) 'A proposed framework for Indian Sign Language Recognition', in International Journal of Computer Application, Vol. 5, pp.158-169.

[10] Nandy, Anup & Mondal, Soumik & Prasad, Jay & Chakraborty, Pavan & Nandi, G. (2010) 'Recognizing & interpreting Indian Sign Language gesture for Human Robot Interaction', International Conference on Computer and Communication Technology (ICCCT), pp.712–717.

[11] Baranwal, Neha & Nandi, G. (2017) 'An efficient gesture based humanoid learning using wavelet descriptor and MFCC techniques', International Journal of Machine Learning and Cybernetics, Vol. 8, pp. 1369 – 1388.

[12] Wazalwar, Sampada & Shrawankar, Urmila. (2017) 'Interpretation of sign language into English using NLP techniques', Journal of Information and Optimization Sciences, Vol. 38, pp. 895–910.

[13] Tripathi, Kumud & Nandi, Neha. (2015) 'Continuous Indian Sign Language Gesture Recognition and Sentence Formation' Procedia Computer Science, Vol. 54, pp.523–531.

[14] Shorten, Connor & Khoshgoftaar, Taghi. (2019) 'A survey on Image Data Augmentation for Deep Learning', Journal of Big Data, Vol. 6, pp. 1–48.

[15] Mikołajczyk, Agnieszka & Grochowski, Michał. (2018) 'Data augmentation for improving deep learning in image classification problem' in International Interdisciplinary PhD Workshop (IIPhDW). pp.117–122.

[16] Wadhawan, Ankita & Kumar, Parteek. (2019) 'Sign Language Recognition Systems: A Decade Systematic Literature Review', Archives of Computational Methods in Engineering, pp. 1-29.

[17] Er-Rady, Adil & Faizi, R. & Rachid, Oulad haj thami & Housni, H. (2017) 'Automatic sign language recognition: A survey', in International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). pp.1–7.

[18] Forster, Jens & Schmidt, Christoph & Hoyoux, Thomas & Koller, Oscar & Zelle, Uwe & Piater, Justus & Ney, Hermann. (2012) 'RWTH-PHOENIX- Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus' in International Conference on Language Resources and Evaluation (LREC).

[19] Forster, Jens & Schmidt, Christoph & Koller, Oscar & Bellgardt, Martin

& Ney, Hermann. (2014) 'Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather' in International Conference on Language Resources and Evaluation (LREC).

[20] Koller, Oscar & Forster, Jens & Ney, Hermann. (2015) 'Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers', Computer Vision and Image Understanding, Vol. 141, pp.108–125.

[21] Thangali, Ashwin & Nash, Joan & Sclaroff, Stan & Neidle, Carol. (2011) 'Exploiting phonological constraints for handshape inference in ASL video' in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.521–528.

[22] Oszust, Mariusz & Wysocki, Marian. (2013) 'Polish sign language words recognition with Kinect' in 6th International Conference on Human System Interactions (HSI), pp. 219-226.

[23] Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., & Rosete, A. (2016) LSA64: An Argentinian Sign Language Dataset.

[24] Chen, Chen & Zhang, Baochang & Hou, Zhenjie & Jiang, Junjun & Liu, Mengyuan & Yang, Yun. (2016) 'Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features', Journal of Multimedia Tools and Applications, Vol. 76, pp.4651–4669.

[25] Adithya, V. & Vinod, P.R. & Gopalakrishnan, Usha. (2013) 'Artificial neural network-based method for Indian sign language recognition' in IEEE Conference on Information & Communication Technologies, pp.1080-1085.

[26] Agrawal, Subhash & Jalal, Anand & Bhatnagar, Charul. (2012) 'Recognition of Indian Sign Language using feature fusion', in 4th International Conference on Intelligent Human Computer Interaction: Advancing Technology for Humanity, IHCI 2012, pp.1–5.

[27] Kumar, Pradeep & Saini, Rajkumar & Behera, Santosh & Dogra, Debi & Roy, Partha. (2017c) 'Real-time recognition of sign language gestures and air-writing using leap motion', in Fifteenth IAPR International Conference on Machine Vision Applications (MVA), pp. 157–160.

[28] Rahaman, Muhammad & Jasim, Mahmood & Ali, Md & Hasanuzzaman, M. (2014) 'Real-time computer vision-based Bengali Sign Language recognition' in 17th International Conference on Computer and Information Technology (ICCIT), pp.192–197.

[29] Uddin, Azher & Chowdhury, Shayhan. (2016) 'Hand sign language recognition for Bangla alphabet using Support Vector Machine' in International Conference on Innovations in Science, Engineering and Technology (ICISET), pp.1–4.

[30] Yasir, Farhad & P.W.C, Prasad & Alsadoon, Abeer & Elchouemi, Amr. (2015) 'SIFT based approach on Bangla sign language recognition', in 8th International Workshop on Computational Intelligence and Applications (IWCIA), pp. 35-39.

[31] Mehrotra, Kapil & Godbole, Atul & Belhe, Swapnil. (2015) 'Indian Sign Language Recognition Using Kinect Sensor' in International Conference Image Analysis and Recognition, pp.528–535. https://doi.org/10.1007/978-3-319-20801-5_59.

[32] Kumar, Pradeep & Gauba, Himaanshu & Roy, Partha & Dogra, Debi. (2017b) 'A Multimodal Framework for Sensor based Sign Language Recognition', Neurocomputing, Vol. 259, pp. 21-38.

[33] Kumar, Pradeep & Gauba, Himaanshu & Roy, Partha & Dogra, Debi. (2017a) Coupled HMM-based multi-sensor data fusion for sign language recognition, Pattern Recognition Letters, Vol. 86, pp.1–8.

[34] Kumar, Pradeep & Saini, Rajkumar & Roy, Partha & Dogra, Debi. (2018) 'A position and rotation invariant framework for sign language recognition (SLR) using Kinect', Multimedia Tools and Applications, Vol. 77, pp. 8823-8846.

[35] Sahoo, Ashok & Sarangi, Pradeepta & Goyal, Parul. (2019) Indian Sign Language Recognition using Soft Computing Techniques, Machine Vision Inspection Systems: Image Processing, Concepts, Methodologies and Applications, Wiley.

[36] Kishore, P.V.V. & Prasad, M.V.D. & Anil Kumar, D. & Sastry, A. (2016) 'Optical Flow Hand Tracking and Active Contour Hand Shape Features for Continuous Sign Language Recognition with Artificial Neural Networks' in 6th International Conference on Advanced Computing (IACC).

[37] Kishore, P.V.V. & Anil Kumar, D. (2016) 'Selfie continuous sign language recognition using neural network' in IEEE Annual India Conference (INDICON), pp. 1–6.

[38] Mariappan, H & Gomathi, V. (2019) 'Real-Time Recognition of Indian Sign Language' in International Conference on Computational Intelligence in Data Science (ICCIDS). pp.1–6.

[39] Bhavsar, Hemina & Trivedi, Jeegar. (2020) 'Indian Sign Language Recognition Using Framework of Skin Color Detection, Viola- Jones Algorithm, Correlation-Coefficient Technique and Distance Based Neuro-Fuzzy Classification Approach', Emerging Technology Trends in Electronics, Communication and Networking, Vol. 1214. pp.235–243.

[40] Lopez-Moreno J. (2015) 'Compositing and Chroma Keying' In Luo R. (eds) Encyclopedia of Color Science and Technology. Springer, Berlin, Heidelberg. pp. 1-8.

[41] Perez, Luis & Wang, Jason. (2017) 'The Effectiveness of Data Augmentation in Image Classification using Deep Learning',ArXiv: abs/1712.04621.

[42] Lecun, Yann & Kavukcuoglu, Koray & Farabet, Clement. (2010) 'Convolutional networks and applications in vision', in proceedings of 2010 IEEE International Symposium on Circuits and Systems, pp. 253-256.

[43] Howard, Andrew & Zhu, Menglong & Chen, Bo & Kalenichenko, Dmitry & Wang, Weijun & Weyand, Tobias & Andreetto, Marco & Adam, Hartwig. (2017) 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications', arXiv:1704.04861

[44] Powers, David. (2008) 'Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation', ArXiv, abs/2010.16061.