Towards Performance Improvement in Indian Sign Language Recognition

Kinjal Mistree
Computer Engineering Department
Uka Tarsadia University
Bardoli
kinjal.mistree@utu.ac.in

Devendra Thakor
Computer Engineering Department
Uka Tarsadia University
Bardoli
devendra.thakor@utu.ac.in

Brijesh Bhatt
Computer Engineering Department
Dharmsinh Desai University
Nadiad
brij.ce@ddu.ac.in

Abstract

Sign language is a complete natural language used by deaf and dumb people. It has its own grammar and it differs with spoken language to a great extent. Since people without hearing and speech impairment lack the knowledge of the sign language, the deaf and dumb people find it difficult to communicate with them. The conception of system that would be able to translate the sign language into text would facilitate understanding of sign language without human interpreter. This paper describes a systematic approach that takes Indian Sign Language (ISL) video as input and converts it into text using frame sequence generator and image augmentation techniques. By incorporating these two concepts, we have increased dataset size and reduced overfitting. It is demonstrated that using simple image manipulation techniques and batch of shifted frames of videos, performance of sign language recognition can be significantly improved. Approach described in this paper achieves 99.57% accuracy on the dynamic gesture dataset of ISL.

1 Introduction

According to (Durkin and Conti-Ramsden, 2010), sign language can be considered as a collection of gestures, movements, postures, and facial expressions corresponding to letters and words in spoken languages. Comparing sign language with spoken language, main difference exists on modality. Deaf and dumb people use sign language as means of communication to express their thoughts and emotions. Each country has its own sign language with high degree of grammatical variations. Indian sign language is largely dominated by object features, and is different than other sign languages.

Trained sign language interpreters are needed during medical and legal appointments, educational and training sessions, to provide interpreting services. Over the past few years, there has been an increasing demand for human interpreters. But in India, approximately 300 certified human interpreters are available (vid). Given a shortage of interpreting services, a system can be designed that offers flexible alternative when human interpreters are not available.

Sign language recognition is an approach that converts input sign gesture(s) into text or speech. Sign language recognition is a very challenging task since this task involves interpretation between visual and linguistic information. Deep neural networks (DNN) perform remarkably well for image recognition task (Shorten and Khoshgoftaar, 2019). But these networks are heavily reliant on big data, otherwise they lead to overfitting. Overfitting refers to the phenomenon when a network learns a function with very high variance such as to perfectly model the training data (Shorten and Khoshgoftaar, 2019). Strategies to avoid overfitting and to increase generalization performance of DNNs are dropout, batch normalization, transfer learning, one-shot learning and pretraining. In contrast to these techniques, data augmentation approaches problem of overfitting and generalization from the root of the problem, the training dataset. We have adopted this concept on input videos for ISL recognition to make dataset inflated.

In order to increase dataset size, we have created sequence of frames in batches systematically. We have proposed this approach to generate more instances for ISL recognition task in order to achieve better model performance. In particular, we have addressed the issue of one research question: how to use DNN 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.

The rest of the paper is organized as follows: In Section 2 we have discussed work related to Indian sign language recognition. Section 3 explains steps of our proposed approach in details. Section 4

Proceedings of the 17th International Conference on Natural Language Processing, pages 349–354
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)

shows dataset description and experimental results along with analysis. Section 5 provides concluding remarks and directions for future work.

2 Related Work

There are two approaches for sign language recognition: 1) Device based approach and 2) Vision based approach. Device based approach hinders the natural movement of hands. Also, signer has to wear gloves that are connected through computer through cables. So device based approaches appear to be complex, costly and difficult to deploy. Considering limitations of device based approach, we have discussed related work with respect to vision based approach in ISL recognition.

(Rekha et al., 2011) proposed an approach that recognized 66 ISL dynamic samples of ISL alphabets. These gestures were classified using Dynamic Time Warping and approach achieved 77.2% accuracy. (Tripathi and Nandi, 2015) presented an approach for recognizing ISL sentences in ISL with 91% accuracy with Hidden Markov Model. Authors recorded single handed and double handed dynamic samples from 10 sentences. Recognition rate of 90.17% was achieved by (Kishore et al., 2016) for 580 samples. Though these samples were not recorded according to ISL grammar, this approach worked for manual components of ISL. (Kumar et al., 2016) used front camera of the mobile phone for collecting dynamic signs. The authors achieved 90% accuracy by extracting hand and head contour energies from the collected dynamic signs.

Issues like hand segmentation from upper half of the body image, boundary changes depending on hand shape of various signers are solved by (Baranwal and Nandi, 2016) with 85% accuracy. Authors have used Otsu thresholding method for segmentation, Mel Sec Frequency Cepstral Coefficients (MFCC) for feature extraction of 8 dynamic samples. 92% accuracy was achieved by (Baranwal et al., 2017) using concept of possibility theory on 20 different videos of continuous gestures. These videos were captured in different backgrounds like black, red, multiple objects with black full sleeves dress. (Wazalwar and Shrawankar, 2017) used pseudo 2-dimensional Hidden Markov Model (P2DHMM) for feature extraction, which was proven better than simple Hidden Markov Model. For converting recognized signs in English text, LALR parser was used and 90% accuracy was achieved. (Muthu Mariappan and Gomathi, 2019) proposed an approach that extracted head and hand contour energies using front camera of the mobile for collecting signs.Approach presented by (Sahoo and Ravulakollu, 2014) used skin color segmentation and artificial neural network (ANN). This approach worked for specific signer as ANN was trained by taking face and hand features of specific signer.

As far as reserach in ISL is concerned, after nearly 35 years of research, ISL is still in infancy when compared to other international sign languages. All the approaches described here work under controlled laboratory setting and use feature extraction techniques. This motivated us to work for ISL recognition by doing generalized settings in very limited amount of input data without decreasing accuracy of ISL recognition.

3 Method

The details of the dataset are described in the Section 4. In this section we describe the steps of the approach we followed to convert ISL video into text. Figure 1 shows the general framework of our approach.

3.1 Video to frame conversion

Each ISL input video is converted in RGB frames. Videos were originally captured at 30 fps, so accordingly frames are generated for each video of different length.

3.2 Horizontal flipping

Signer is either left-handed or right-handed. On the batch of frames, horizonal flipping is performed in order to incorporate both left-handed and right-handed signs.

3.3 Frame sequence generator with data augmentation

To identify main frames, one possibility is that we pick N distributed frames 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, we have created a video generator that provides decomposed frames for the entire video in sequential form to get more sequences from one video file. For each 30 FPS video we have used this video generator to select 5 frames per second. We have decided to select every 6th frame based on analysis of histogram difference in frames. For each individual video, frames are

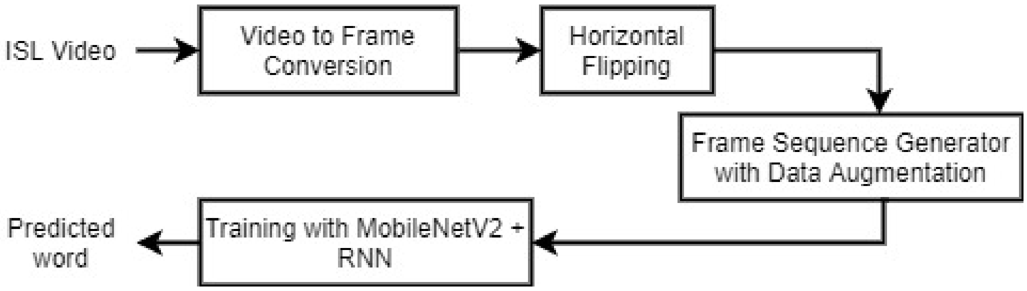
ISL Video
Video to Frame Conversion
Horizontal Flipping
Frame Sequence Generator with Data Augmentation
Training with MobileNetV2 + RNN
Predicted word

Figure 1: General framework of frame sequence generator based approach

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.

(Perez and Wang, 2017) have discussed how to produce promising ways to increase the accuracy of classification tasks using data augmentation. We have 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. We have 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.

3.4 Training with MbileNetV2 + RNN

Image augmentation increases the size of the dataset which is originally very small but the data similarity is still very high. Transfer learning works well with limited and similar data samples by transferring knowledge from models pretrained on large datasets. Among the popular pretrained models, we have used MobileNetV2 as it is light-weight, low-latency deep neural network best suited with restricted resources in mobile and embedded vision applications (Sandler et al., 2018). We have empirically changed the configuration of the top layers of the MobileNetV2 model in order to get the best recognition accuracy. Based on this, top 9 layers of the model are selected for retraining with the augmented frame sequence. This is injected in one time-distributed layer at the end to have the one-dimensional shape compatible with LSTM layer. Finally, dense layer is added to get the prediction of ISL word.

4 Experimental Results and Analysis

In this section, we present the details of the dataset and the experimental results with analysis.

4.1 Dataset

(Nandy et al., 2010) created repository of static and dynamic hand gestures of 21 specific kind of ISL words under various light illumination conditions. Out of 21 classes, 11 classes corresponds to dynamic hand gestures and 10 classes corresponds to static hand gestures. The dataset was created in July 2009 at the Robotics and AI Lab, IIIT-Allahabad, having frame resolution of 320 * 240 pixels. Statistics of training, validation and testing samples used in work are shown in Table 1.

4.2 Results and analysis

Training and testing data used by us was prepared under various light illumination conditions but with identical camera settings. As discussed in previous section, we have chosen parameter range in order to incorporate randomness in sample generation. We have excluded horizontal flipping from this set of augmentation techniques because we already inflated dataset in order to incorporate both left-handed as well as right-handed signs. Table 2 shows type of augmentation techniques and parameter range used for our experiments.

Parameters	Values
No. of classes (ISL words)	10
Training samples	79
Validation samples	15
Testing samples	27
Training samples after horizontal flipping	158
Validation samples after horizontal flipping	30
Testing samples after horizontal flipping	54
No. of training sequences after using frame sequence generator	12692
No. of validation sequences after using frame sequence generator	2443
No. of testing sequences after using frame sequence generator	4082

Table 1: Image sequences after using frame sequence generator with data augmentation

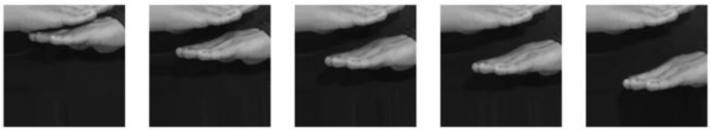

Figure 2: Sample frames 5, 11, 17, 23 and 29 for sign ’below’ as per selection by frame sequence generator. This frame set is result of frame sequence generator + image augmentation techniques.

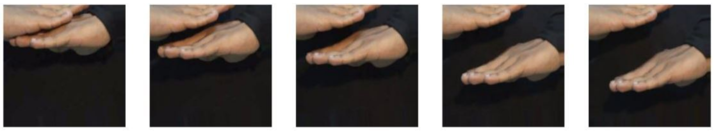

Figure 3: Another set of frames 2, 8, 14, 20 and 26 for sign ’below’ as result of frame sequence generator + image augmentation techniques.

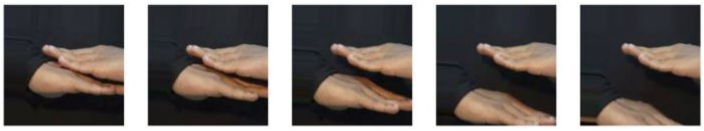

Figure 4: Sample frames 3, 9, 15, 21 and 27 for sign ’below’ as per selection by frame sequence generator. This frame set is result of horizonal flipping + frame sequence generator + image augmentation techniques.

Augmentation Type	Parameter Range
Zooming	[0.7, 1]
Rotation	[0, 15]
Vertical shifting	[0, 0.3]
Horizontal shifting	[0, 0.3]
Brightness	[0.2, 1.0]
Shear angle	[0, 0.3]

Table 2: Augmentation techniques and range of parameters used for each ISL input video

Figure 2, 3 and 4 shows various frame sequences generated by video generator for sign ’below’, using frame sequence generator.

By following the steps explained in the previous section, an experiment was conducted on ISL dynamic gesture dataset for 10 categories of ISL words. Table 3 shows model performance on ISL dynamic gesture dataset using MobilNetV2 + RNN, by keeping top 6, 9 and 12 layers trainable, while keeping other layers of model frozen. We have achieved recognition accuracy as 99.57% when keeping last 9 layers trainable, which outperforms the overall accuracy reported by (Nandy et al., 2010).

Figure 5 and Figure 6 shows plot of accuracy and

MobileNetV2+RNN	Trainable layers = 6	Trainable layers = 9	Trainable layers = 12
Training accuracy	98.33%	99.41%	93.12%
Validation accuracy	98.67%	100%	93.85%
Testing accuracy	93.91%	99.57%	94.15%
Training loss	0.2134	0.0313	0.3874
Validation loss	0.0976	0.0010	0.2916
Testing loss	0.0841	0.0016	0.0287

Table 3: Model performance on ISL dynamic gesture dataset

loss, by keeping last 9 and layers as trainable. We have trained model for 40 epochs and used early stopping on validation loss, so the training gets stopped when there is no significant improvement in accuracy after 3 continuous epochs.

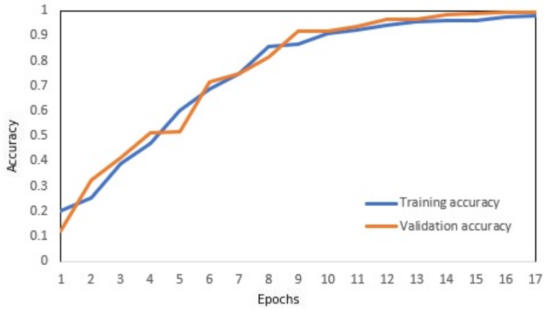

Figure 5: Training and validation accuracy by keeping last 9 layers trainable using MobileNetV2+RNN

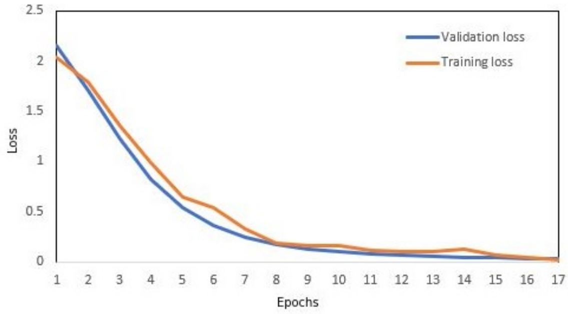

Figure 6: Training and validation loss by keeping last 9 layers trainable using MobileNetV2+RNN

Approach	Accuracy
(Nandy et al., 2010)	81.94%
Our method	99.57%

Table 4: Accuracy of proposed approach on ISL dynamic gesture dataset

Table 4 shows comparison of recognition result in terms of accuracy, for approach presented by (Nandy et al., 2010) and our approach. In previous work, results for 11 ISL words are presented. We have excluded result of class ’Yes’ as dataset provided by authors has not sufficient samples for sign ’Yes’. Also, we have used less number of training, validation and testing samples in order to evidently prove the effect of our proposed approach in classification.

5 Conclusion and Future work

It becomes a challenging task when we want to achieve more accuracy with less 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. Image augmentation also increases accuracy of test data. In this work, we have 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. We have proposed an approach that uses pretrained model MobileNetV2 to learn features from augmented frame sequences of ISL gestures using batch of shifted frames to provide decayed sequences for the same gesture.

We are working on extending our work from lexical level analysis to machine translation to generate ISL sentences. We are also in the process of creating new dataset of ISL sentences using the dictionary launched by Indian Sign Language Research and Training Centre (ISLRTC). In future, we will compare results of MobileNetV2 model with other pretrained models on our dataset.

References

www.islrtc.nic.in/history-0. Accessed April 28, 2019.

Neha Baranwal and G. Nandi. 2016. An efficient gesture based humanoid learning using wavelet descriptor and mfcc techniques. *International Journal of Machine Learning and Cybernetics*, 8.

Neha Baranwal, Avinash Singh, and G. Nandi. 2017. Development of a framework for humanrobot interactions with indian sign language using possibility theory. *International Journal of Social Robotics*, 9.

Kevin Durkin and Gina Conti-Ramsden. 2010. Young people with specific language impairment: A review of social and emotional functioning in adolescence. *Child Language Teaching & Therapy - CHILD LANG TEACH THER*, 26:105–121.

P.V.V. Kishore, M.V.D. Prasad, D. Anil Kumar, and A Sastry. 2016. Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks.

D. A. Kumar, P. V. V. Kishore, A. S. C. S. Sastry, and P. R. G. Swamy. 2016. Selfie continuous sign language recognition using neural network. In *2016 IEEE Annual India Conference (INDICON)*, pages 1–6.

H. Muthu Mariappan and V. Gomathi. 2019. Real-time recognition of indian sign language. In *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pages 1–6.

Anup Nandy, Jay Prasad, Soumik Mondal, Pavan Chakraborty, and G. Nandi. 2010. Recognition of isolated indian sign language gesture in real time. volume 70, pages 102–107.

Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning.

J. Rekha, J. Bhattacharya, and S. Majumder. 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)*, pages 30–35.

Ashok Sahoo and Kiran Ravulakollu. 2014. Indian sign language recognition using skin colour detection. *International Journal of Applied Engineering Research*, 9:7347–7360.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48.

Kumud Tripathi and Neha Nandi. 2015. Continuous indian sign language gesture recognition and sentence formation. *Procedia Computer Science*, 54:523–531.

Sampada Wazalwar and Urmila Shrawankar. 2017. Interpretation of sign language into english using nlp techniques. *Journal of Information and Optimization Sciences*, 38:895–910.