Global Journal of Computer Science and Technology: D
Neural & Artificial Intelligence
Volume 21 Issue 2 Version 1.0 Year 2021
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Sign Language Recognition for Static and Dynamic Gestures

By Jay Suthar, Devansh Parikh, Tanya Sharma & Avi Patel

Abstract- Humans are called social animals, which makes communication a very important part of humans. Humans use shoes and non-verbal forms of language for communication purposes, but not all humans can give oral speech. Hearing impaired and mute people. Sign language became consequently advanced for them and nevertheless impairs communication. Therefore, this paper proposes a system that uses streams to use CNN networks for the classification of alphabets and numbers. Alphabet and number gestures are static gestures in Indian sign language, and CNN is used because it provides very good results for image classification. Use hand-masked (skin segmented) images for model training. For dynamic hand gestures, the system uses the LSTM network for classification tasks. LSTMs are known for their accurate prediction of time zone distributed data. This paper presents different types of hand gestures, namely two models for static and dynamic prediction, CNN and LSTM.

Keywords: indian sign language, skin-segmentation, CNN (convolutional neural network), LSTM (long shortterm memory).

GJCST-D Classification: I.2.7

Strictly as per the compliance and regulations of:

Sign Language Recognition for Static and Dynamic Gestures

Jay Suthar [α], Devansh Parikh [σ], Tanya Sharma [ρ] & Avi Patel [ω]

Abstract- **Humans are called social animals, which makes communication a very important part of humans. Humans use shoes and non-verbal forms of language for communication purposes, but not all humans can give oral speech. Hearing impaired and mute people. Sign language became consequently advanced for them and nevertheless impairs communication. Therefore, this paper proposes a system that uses streams to use CNN networks for the classification of alphabets and numbers. Alphabet and number gestures are static gestures in Indian sign language, and CNN is used because it provides very good results for image classification. Use hand-masked (skin segmented) images for model training. For dynamic hand gestures, the system uses the LSTM network for classification tasks. LSTMs are known for their accurate prediction of time zone distributed data. This paper presents different types of hand gestures, namely two models for static and dynamic prediction, CNN and LSTM.**

Keywords: indian sign language, skin-segmentation, CNN (convolutional neural network), LSTM (long short-term memory).

Nomenclature:
ISL: Indian Sign Language
ROI: Region of Interest
CNN: Convolutional Neural Network
RMSProp: Root Mean Square Propagation
LSTM: Long Short-Term Memory

I. Introduction

Everyone uses language to communicate with others, whether it is English, Spanish, sign language, or touch or smell language. Sign language is the language used by deaf-mute people to talk. It varies from country to country and has its own vocabulary. Indian Sign Language (ISL) is a collection of gestures used by the deaf community in India. These gestures are also different in different parts of India.

It is always a challenge for normal people to communicate with deaf-mute people and vice versa. Sign language translation is the solution to this problem. It provides a bridge of communication between the community at large and the deaf-mute community. There are two main methods for the recognition of sign language, glove-based and computer vision-based [1]. In this article, a computer vision-based approach to interpreting ISL in two different ways is discussed. ISL letter recognition includes camera frame extraction, hand masking, feature extraction and classification recognition. This is to identify the alphabet from a single frame. The second method is to recognize gestures through words. The camera frame sequence is used to recognize gestures. It consists of the same modules as letter recognition, but it uses a series of frames instead of one frame. This article focuses on ISL recognition through deep learning and computer vision. The rest of thesis is organized as follows; the second part presents the related work done in gesture recognition. The third part contains the methodology of the two methods of recognition of the Indian sign language. The first method is suitable for static gestures and the second method is suitable for dynamic gestures. Discussion of the results and conclusions are explained in Sections 4 and 5, respectively.

II. Related Work

Many techniques have been developed to recognize sign language. There are two main approaches that use tracking sensors or computer vision to track various movements. Much research has been done on sensor-based approaches using gloves and wires [1, 2, 3]. Therefore, it is inconvenient to wear these devices continuously. Additional work will primarily focus on computer vision-based approaches.

A lot of work has been done using a computer vision based approach. The authors have proposed various methods of recognizing sign language using CNN (Convolution Neural Network), HMM (Hidden Markov Model) and contour lines [4, 5, 6, 7]. Different methods are used to split images such as HSV and color difference images [4, 5]. The authors proposed an SVM (Support Vector Machine) method for classification [6, 8]. Archana and Gajanan also compared different methods for partitioning and feature extraction [9]. All previous treatises have successfully recognized the ISL alphabet. But in reality, deaf or mute people use speech gestures to convey messages. If the word has a static action, you can use these previous methods to check the word.

ISL Many words required hand movements. The image classification method is not a simple image classification technique, but is suitable for identifying these dynamic gestures. Video-based action recognition has already attracted attention in several studies [10, 11, 12]. Instead of capturing the color image data for each

Author α σ ρ ω: e-mails: jaysuthar0603@gmail.com, devansh.parikh98@gmail.com, tanyasharma4869@gmail.com, avipatel2013@gmail.com

frame of the video, some researchers performed differences between successive frames and randomly provided these segments to the TSN (Temporal Segment Network) [11]. Sun, Wang and Yeh use LSTM (Long Short-Term Memory) [13] to describe video classification and captions. Juilee, Ankita, Kaustubh and Ruhina used video to suggest a method of hydration recognition in India [14]. As a result of searching the sign language recognition system, most studies use static sign language gestures and video recognition techniques to study dynamic gesture identification and only perform video classification for various actions discovered.

III. Methodology

a) Static gesture classification

Experiments were performed on the data set provided by [15]. The dataset contains 36 folders representing 09 and A-Z, each consisting of an image of a hand subdivided by the corresponding alphanumeric skin color. There are 220 images of 110 x 110 pixels each for each alphanumeric character. Figure 1 shows an image of each label in the data set.

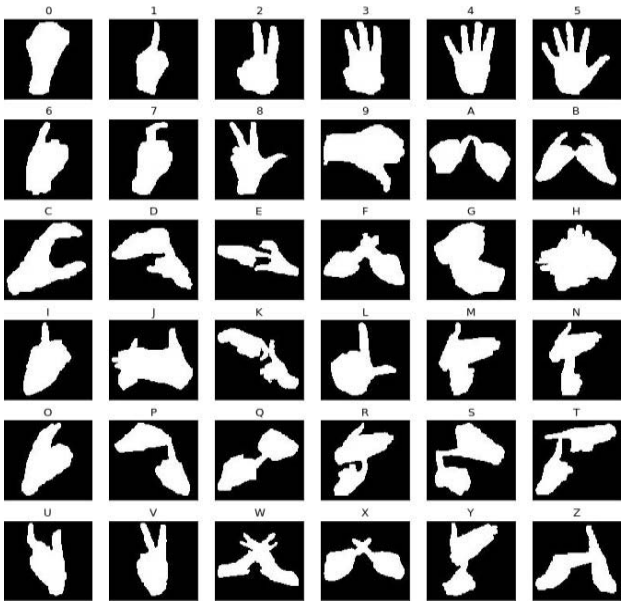

Figure 1: Hand-masked image data set

This image is split into a training dataset and a test dataset all in a ratio of 80:20. So the training dataset contains 6336 images and the test dataset contains 1584 images corresponding to 36 classes. Also, since the number of images per class is small, we performed data expansion to feed more data to the CNN model. Data scaling includes operations such as rotation, width_shift, height_shift, rescaling, etc. The specification of the CNN model is shown in Table 1 below.

Table 1: CNN Specification

Property	Value
ConvolutionLayer	3Layers (32,64,128 nodes)
ConvolutionLayer(KernelSize)	3,3, 2
MaxPoolingLayer	3 Layers-(2, 2)
FullyConnected Layer	128nodes
OutputLayer	36nodes
ActivationUsed	Softmax
Optimizer	RMSProp
Hyperparameters	
Learningrate	0.01
No.ofepochs	10

After training the model, predict the output by performing the following steps:

Frame Extraction: Uses OpenCV library to capture video from webcam for live prediction. After capturing the

video, take a single frame and define a region of interest (ROI) in that frame. The area of interest is the area in which a person runs a stream.

Skinsegmentation: The ROI of the frame is transformed into a hand-masked image to provide to the model for predictive purposes. First, you need to blur the image to reduce noise. This is done by applying Gaussian Blur. After blurring ROI is converted to HSV color scale in RGB. Converting an image to the HSV color scale helps detect better skin than RGB. Next, lower and upper limits are set for skin extraction. Here, (108, 23, 82) was used in the low range and (179, 255, 255) was used in the high range. This range offers us the best results. After selecting a range, compare the values of each pixel and if the pixel value is not within the range, it will be converted to black, otherwise it will be converted to white pixels. This provides us with handmasked images. Still, the hand-masked image is noisy and the edges are not aligned. To solve this problem, use the Dilate and Erode features available in OpenCV to smooth the edges.

Prediction: The ROI of the frame is converted to a handmasked image. This hand-masked image is provided as input to the CNN model for prediction. 09 or AZ is provided for the output of the original frame, but it is a predicted value. But this leads to another problem, the frame output keeps blinking. To solve this problem, we used a 25-frame forecast and used the maximum forecast class as the output.

Figure 2 shows a handmasked image of the alphabet L and the final output.

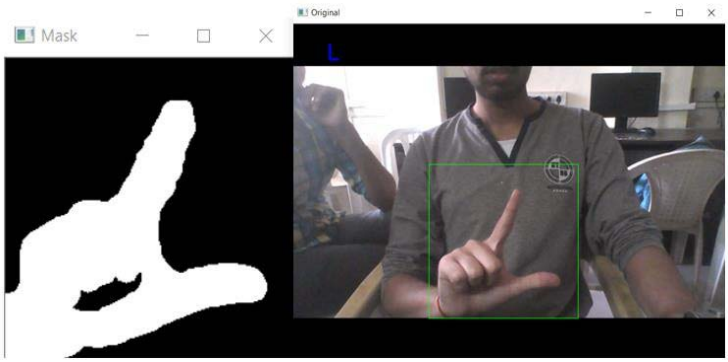

Figure 2: Hand-mask and Predicted Output

b) Dynamic gesture classification

Neural networks can help predict complex data and values most of the time. Input is not related to time or is not required in chronological order. This is the case for static gestures in ISL, so a multi-layer CNN architecture is sufficient. However, for dynamic gestures, you cannot perform CNN silver because you have to keep the previous state. Therefore, LSTM networks are useful in this case. LSTM is an RNN (Recurrent Neural Network) type that has a structure similar to a chain of repeating modules that is useful for learning long-term dependencies from sequential data.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
time_distributed_1 (TimeDist (None, 8, 7, 7, 1280)     2257984
_________________________________________________________________
time_distributed_2 (TimeDist (None, 8, 1280)           0
_________________________________________________________________
lstm_1 (LSTM)                (None, 8, 64)             344320
_________________________________________________________________
dropout_1 (Dropout)          (None, 8, 64)             0
_________________________________________________________________
lstm_2 (LSTM)                (None, 64)                33024
_________________________________________________________________
dense_1 (Dense)              (None, 64)                4160
_________________________________________________________________
dropout_2 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 24)                1560
_________________________________________________________________
dropout_3 (Dropout)          (None, 24)                0
_________________________________________________________________
dense_3 (Dense)              (None, 5)                 125
=================================================================
Total params: 2,641,173
Trainable params: 2,607,061
Non-trainable params: 34,112
```

Figure 3: Model Architecture

First, to train the model, we need the data. I have created a dataset that contains 12 classes of video. Today, tomorrow, yesterday, goodbye, mom, dad, time, eat, well, I (I), thank you, Namaste class. Each class has about. 57 seconds 2025 hand gesture video. The video is captured using the phone back

camera at an average of 2830 frames per second. Data sets are created among different people from different backgrounds so that different situations can be trained. Currently this dataset is split into 20ta for validation and the rest for training. The model structure is shown in Figure 3. The input continuously delivers a sequence of 8 frames/images extracted from images in the training dataset. Apply an RGB difference filter before serving these 8 frames as input. The RGB difference subtracts the current frame from the previous frame. Therefore, only the changed pixels remain in the frame and the remaining still images are deleted. In this way, it helps to capture time-varying visual features. Here in our case it helps to capture the gesture pattern and the background is also removed so it becomes independent in a variety of background scenarios.

If so, these frame sizes change to 224 x 224 pixels. This is because the next layer is a MobileNetV2 layer that only accepts image sizes up to 224 x 224 pixels. As a pre-trained model, we used the weight of `Imagenet` and MobileNetV2. MobileNetV2 can be used as a pre-trained model used for image segmentation, eliminating the task of building CNN models for image segmentation. Separate the Mobile Net V2 layer for passing these 8 frames with Time Distributed layer used. Here I use the Time Distributed Global Average Pooling layer as I need to flatten the frames to insert a series of frames into the LSTM. Finally, there is a multi-layer LSTM structure with several dropouts and a fully connected layer to reduce the sum of overcharges. LSTMs help recognize pattern formation with dynamic / moving hand gestures. Finally, SGD is used in optimization programs because it provides better results when the available data set is low. Adam provides good results even when the dataset is large.

IV. Results and Discussion

a) Static gesture classification

During checking out of the static hand gesture version and figuring out the greatest architecture, 10epoch models were each trained using various optimizations such as RMSProp, SGD, Adam, etc. By the way, RMSProp gave the best results with a precision of 73.6°. A graph of accuracy and time is shown in Figure 4

Skin segmentation is an integral part of a system for predicting static hand gestures. It was concluded that the lower range (108,23,82) and the higher range (179,255,255) would give the best results. Figure 5-6 shows the gestures predicted to be skin segmentation.

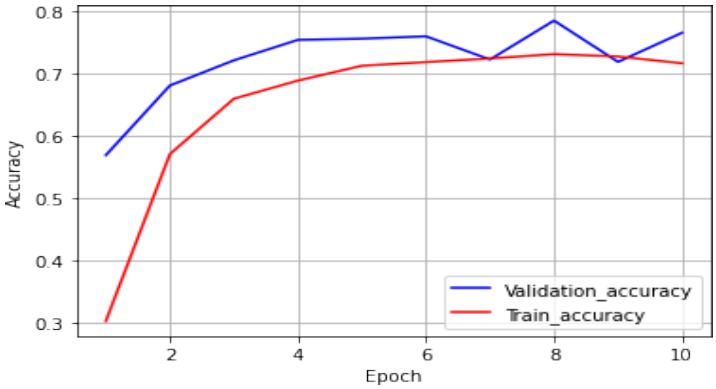

Figure 4: Accuracyvs. Epoch graph for the CNN model

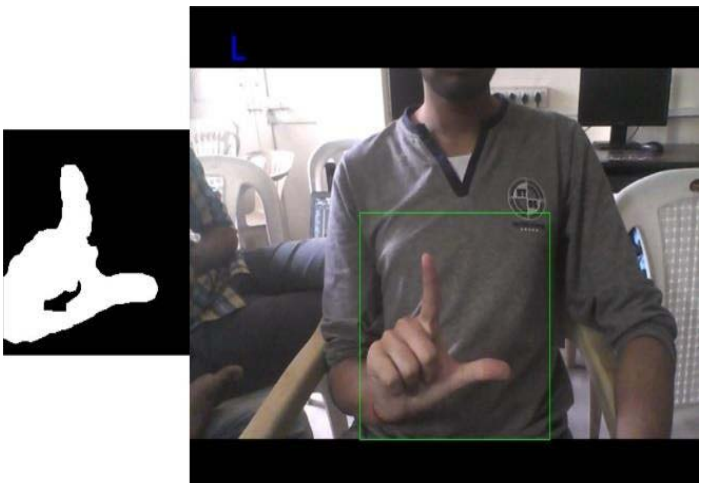

Figure 5: Prediction of Alphabet Letter 'L'

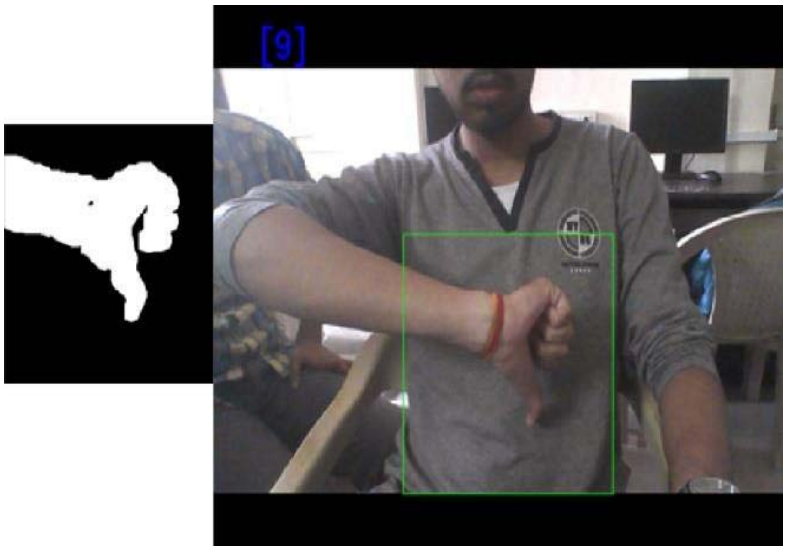

Figure 6: Prediction of Number '9'

There are some limitations to using skin segmentation to recognize static hand gestures. Most importantly, you need a skin-free background. Predictions are wrong because the background contains colors in the skin color range and it is difficult to hide the skin. For example, if the background is a shade of yellow that falls within our range, this problem will occur. The second problem is a stream of similar shape. The equal gestures with alphabets and numbers overlap. For example, the alphabet "V" and the number "2" have the same gesture and cannot be properly distinguished by the system. There is also a similar hand movement problem that reduces accuracy. For example, the letters 'M' and 'N' are very similar. Other similar pairs are `FX` and `1I`.

b) Dynamic gesture classification

Dynamic hand gesture recognition used multi-layered LSTMs to capture patterns formed by movement motions. A model was trained to recognize 12 frequently used words. Model gives medicine. 85 Deputy Pastor of the Diocese. Figures 7 and 8 show the expected behavior and accuracy. If no action is taken, the result is "No action taken".

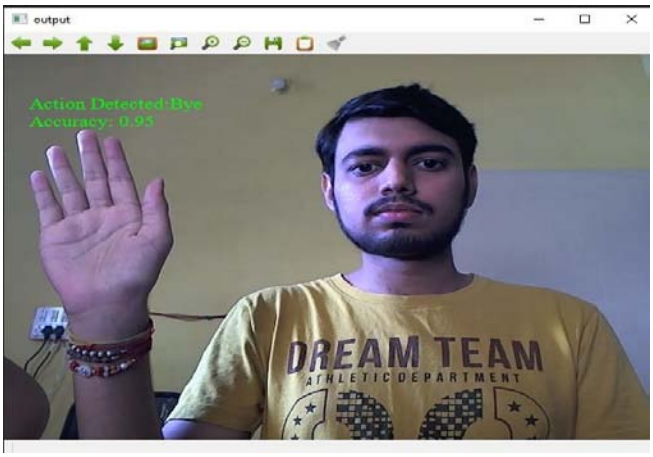

Figure 7: Prediction of Word "Bye"

Figure 8: Prediction of Word "Eat"

Removed static parts of the frame sequence using RGB differences to overcome the background color issue. It also leaves the moving hand in the frame, which helps detect hand gesture patterns. The only problem with this approach is that if the background is moving, the sequence of frames will also have a background, which will affect the prediction accuracy. You can also add more videos to different backgrounds and people's datasets for greater accuracy.

V. Conclusion and Destiny Scope

The Deafmute community is faced with communication challenges every day. This white paper describes two methods for recognizing hand gestures: static gestures and dynamic gestures. For static gesture classification, a CNN model is implemented that classifies the motions alphabetically (AZ) and numerically (09) with a precision of 73. Use hand mask

skin subdivision with the model For dynamic gestures, we trained a model using multi-layer LSTM using 12-word MobileNetV2 and gave very satisfactory results with an accuracy of 85°. For future work with static gestures, another approach to skin segmentation that does not rely on skin color can be built. For dynamic gestures, you can increase the size of data sets with different backgrounds.

References Références Referencias

1. M. Mohandes, M. Deriche, J. Liu, "Image-Based and Sensor-Based Approaches to Arabic Sign Language Recognition", IEEE Transactions on Human-Machine Systems, Vol.44, Issue. 4, pp. 551-557, 2014.
2. C. Zhu, W. Sheng, "Wearable Sensor-Based Hand Gesture and Daily Activity Recognition for Robot-Assisted Living", IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, Vol.41, Issue. 3, pp. 569-573, 2011.
3. T. Jaya, and V. Rajendran, "Hand-Talk Assistive Technology for the Dumb", International Journal of Scientific Research in Network Security and Communication (IJSRNSC), Vol.6, Issue.5, pp.27-31, 2018.
4. L.K.Ramkumar, S.Premchand, G.K.Vijayakumar, "Sign Language Recognition using Depth Dataand CNN", SSRG International Journal of Computer Sciences and Engineering (SSRG- IJCSE), Vol.6, Issue.1, pp. 9-14, 2019.
5. P. Gupta, A. K. Agrawal, S. Fatima, "Sign Language Problem and Solutions for Deaf and Dumb People", In the Proceedings of the International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India,2014.
6. A.S. Nikam, A.G. Ambekar, "Sign Language Recognition Using Image Based Hand Gesture Recognition Techniques", In the Proceedings of the Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, India, pp. 1-5,2016.
7. N. S. Lele, "Image Classification Using Convolutional Neural Network", International Journal of Scientific Research in Computer Science and Engineering (IJSRCSE), Vol.6, Issue.3, pp. 22-26, 2018.
8. J.L.Raheja, A.Mishra and A.Chaudhary, "Indian Sign Language Recognition Using SVM", Pattern Recognition and Image Analysis, Vol.26, Issue. 2, pp. 434-441, 2016.

A. S. Ghotkar, G. K. Kharate, "Study of Vision Based Hand Gesture Recognition Using Indian Sign Language", International Journal on Smart Sensing and Intelligent Systems, Vol.7, Issue.1,2014.

9. K.Simonyan, A. Zisserman, "Two-stream Convolutional Networks for Action Recognition in Videos", Advances in Neural Information Processing Systems, pp. 568-576, 2014.
10. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D Lin, X. Tang, and L.VanGool," Temporal Segment Networks: Towards Good Practices for Deep Action Recognition", In the Proceedings of the European conference on Computer Vision, pp. 20-36. Springer, Cham, 2016.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L.Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks", In the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1725-1732, 2014.

11. J. Sun, J. Wang, T. C. Yeh, "Video understanding: from video classification to captioning". In the Proceedings of the Computer Vision and Pattern Recognition, pp. 1-9, Stanford University, 2017.
12. J. Rege, A. Naikdalal, K. Nagar, R. Karani, "Interpretation of Indian Sign Language through Video Streaming", International Journal of Computer Science and Engineering (IJCSE), Vol. 3, Issue. 11, pp. 58-62, 2015.
13. Pradip Patel, Narendra Patel, "Vision Based Real-time Recognition of Hand Gestures for Indian Sign Language using Histogram of Oriented Gradients Features", in International Journal of Next-Generation Computing, Vol. 10, No. 2, July 2019.