Real-Time Hand Gesture Recognition for Indian Sign Language using CNN with Deep Single-Shot Detector

Barathi Subramanian, Bekhzod Olimov, Jeonghong Kim*
Dept. of Computer Science and Engineering, Kyungpook National University
e-mail: jhk@knu.ac.kr

Abstract: Sign language is a language that uses hand gestures, motions, and facial expression. It primarily used by the deaf and people with difficulties of hearing. However, in real-life, there is always a communication gap between a normal persons and deaf people. Normal people are unable to express and interact well with people having hearing impairment without the knowledge of sign language and it leads to improper interpretation of the signs made by them. To overcome this issue, we have developed a real-time hand gesture recognition model for Indian Sign Language using a Convolutional Neural Network (CNN) with Deep Single-Shot Detector (DSSD). In the conducted experiments, the proposed method achieved approximately 99.5% accuracy on the hand gestures. Also, the proposed method showed its computational efficiency allowing to use the model in real-time applications.

1. Introduction

Sign language is a visually interactive language used by people who are deaf or have difficulties in hearing. Any form of effective should be understandable for both the informant and the recipient. In real life, there is always a communication barrier between the normal hearing community and the people with hearing impairments. Like spoken languages, sign language differs from one geographic region to others; therefore, it is hard for many individuals to understand gesture-based communication that leads to misunderstanding and misinterpretation between people[1]. These issues can also cause more serious problems like learning difficulties, social isolation, and depression for the individuals with hearing disabilities.

There are approximately 300 unique kinds of communication through gestures used around the globe today such as Indian Sign Language (ISL), American Sign Language (ASL), Japanese Sign Language (JSL), Korean Sign Language (KSL), Chinese sign language (CSL) and so on[2]. Indian Sign Language (ISL) is considered to be one of the most complex language compared with the other existing sign languages because of the following reasons[3]:

- It uses both single and double hands for portraying gestures.
- It includes both static and dynamic gestures.
- It uses similar gestures to represent different things.

Considering the complex nature of ISL, we propose a system that alleviates the problem of communication between the normal and the hard of hearing by introducing a CNN-based model for recognizing the hand sign gestures and display its appropriate text on the screen. The proposed method helps to understand the signs (performed by the deaf people) for the normal people without having prior knowledge about the sign language.

In the proposed method, we used CNN techniques, as it is considered to be exceptionally efficient in handling computer vision tasks. Also, it fits for extracting desired features with a high-level accuracy upon adequate training. Since we created our dataset by capturing images in real-time, we can only create a small dataset, which will be insufficient for training a CNN model that usually employs a large dataset for training. In addition, small dataset can cause overfitting. To avoid this issue, we used transfer learning technique for high accuracy prediction using pre-trained models as a starting point, which froze the top layer of the CNN and trained only the last few layers which made a prediction. To identify the region and then to detect the hand position, we implemented DSSD that used a fully convolutional approach, where the network could detect various objects in one single shot with good accuracy that was significantly faster

compared with two-shot regional proposal network (RPN) based approaches.

The rest of the paper is organized as follows. Section 2 discusses currently available methods for sign recognition. Section 3 contains detailed explanation of the proposed methodology. Section 4 represents experimental results. Finally, Section 5 concludes this study with the conclusion and future work.

2. Related Work

Previous studies have emphasized the prediction of sign language gestures in support of people having hearing impairments using the advancement of technologies with Artificial Intelligence algorithms. This literature review highlights some significant works to address the problem in recognition of sign language gestures. Also, we have identified the improvement areas from these works and implemented them in this manuscript. As explained in the introduction, there are numerous sign languages depending on the geographic location. In [4], researchers have proposed a model for Arabic Sign Language (ArSL) translation with a customized glove having 5 flex and one accelerometer sensors. They have also developed an android mobile application to recognize ArSL gestures and translate them to text and audio. But they have translated signs only for letters. In [5], the authors have used a Multilayered Random Forest (MLRF) classifier for sign language recognition with a consumer depth camera. The usage of MLRF contributed to less training time and less memory allocation, but their model lacked the generalization abilities when compared to the other similar methods.

In addition, extensive literature to train CNN for continuous and weakly labelled data was presented [6]. Authors have trained CNN inside an iterative Expectation-Maximization algorithm over 1 million poorly labelled hand shape images denoting the sign language. However, they have achieved only moderate prediction accuracy despite using only the pre-recorded pictures and videos. This might not be suitable for real-time hand shape recognition.

Other literature has explained gesture recognition for ISL [7]. The authors have used Microsoft kinetic sensor to capture the sign images with Support Vector Machine (SVM) and achieved 71.85% recognition accuracy on an average and concluded 100% accuracy for one numeric value and 6 letters alone. However, their work did not consider every context of the gestures. Likewise, there are several other pieces of literature where gesture recognition was performed using machine learning, deep learning methods [8]- [10]. In a review article [11], the authors have discussed about two broad categories based on the approach used for gesture recognition with machine learning algorithms. One is a sensor-based approach and the other is a vision-based approach. From the study, the authors have proposed a technique named sub-unit sign modelling for continuous sign language, which is relating to dynamic gestures. However, there is a lack of information about the kind of data and environment suitable for this technique being discussed in that article.

Considering the aforementioned issues, we propose a real-time gesture recognition system for ISL using transfer learning with CNN, and DSSD.

3. The Proposed Methodology

3.1. Dataset Description:

We prepared a real-time dataset consisting of images with the size of 640 x 480 using a web camera. To recognize a single word, we captured 15 images in various angles and positions to train the model. We divided the image data in a ratio of 70:30 to create the training and testing datasets, respectively.

3.2. Methodology:

After obtaining the training and testing images, we have formulated a CNN model. General overview of the model is shown in Figure 1.

3.3. Proposed Architecture:

There are 4 layers in the proposed layout, and it follows a top-down approach as follows.

1. ISL Data Input Layer
2. Data Preprocessing Layer
3. Data Training & Testing Layer
4. Output Layer

ISL Data Input Layer: This is the input layer that consists of the real-time images captured using a web-camera for each Indian Sign recognition as described in section 3.1.

Data Preprocessing Layer: In this layer, first we perform the process of cropping so that the focus on the hand gesture is not lost due to the unrelated background data. To crop the hands from the image, we have used a graphical annotation tool called labelImg and draw a bounding box

which focuses only on the hands and label the image for its corresponding sign. Then we have used ssd_model_v2_coco model dataset along with our real-time dataset for training. Also, we performed resizing according to the configuration of SSD images that has a fixed dimension of 300 x 300 to avoid misinterpretation of the sign's meaning thereby increasing the likelihood of displaying correct text for that sign. Next the captured images are prone to have noise which brings a necessity to use a filter to smooth the image. We have used Gaussian filter to remove the noise since it performed best in the conducted experiments.

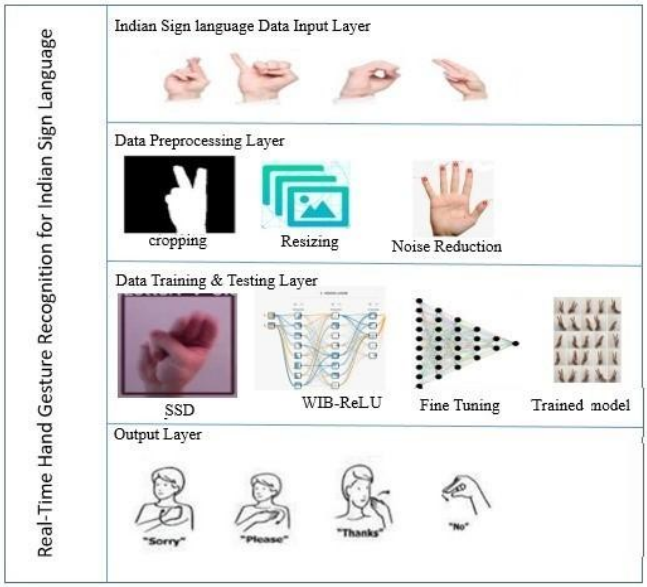


Figure 1. General overview of the proposed method.

Data Training & Testing Layer: Filtered image data is inputted into the model using the concept of CNN with DSSD as shown in the Figure 2 and weight initialization-based rectified linear unit (WIB-ReLU) activation function [12] to create the training and testing model run constructively. High accuracy of recognition is achieved using DSSD to identify the hand region in the image and detect the exact gesture of the hand sign. The activation function helped the model to have good convergence speed. Upon unfreezing a few parts of the top layers in the model base, and jointly training the newly added classifier layers and last layers of the base model, we fine-tune our model to obtain a proficient trained model for inference.

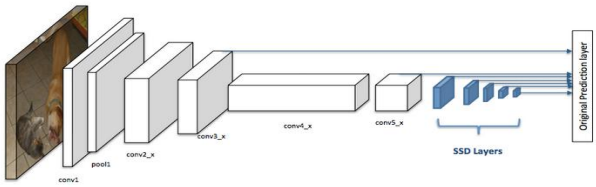


Figure 2. The proposed architecture of a CNN with SSD.

Output Layer: This layer is the text displaying layer, which displays the meaning of the hand gesture in simple text so that a common person can be able to understand the gesture shown.

4.Experimental Results

For implementing the proposed model, we used Python 3.8.4 version on a personal computer with 32GB RAM and Intel i7 3.60GHz CPU, Windows 10 of 64-bit operating system. Also, we used a webcam with resolution of 720p/30fps to capture real-time images as input to the model. Based on the experiment results, the proposed method achieved great prediction accuracy of 99.5 % on average for ten different sign gestures with simple, cost-effective hardware and software requirements. Random hand gesture recognition input images and their corresponding output images for three signs are illustrated in Figure 3.

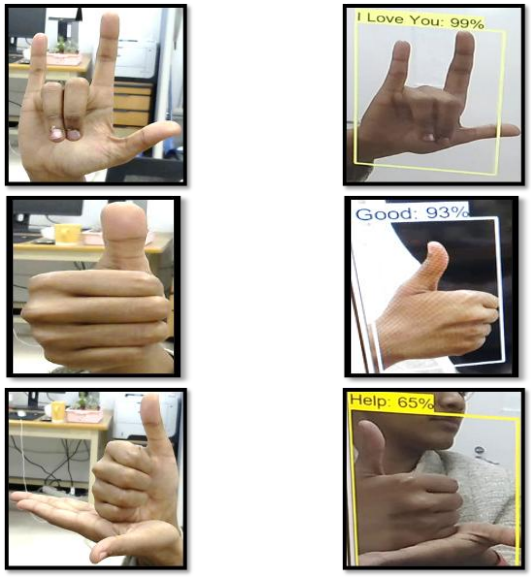


Figure 3. (a)input images into the proposed method (b)output images from the proposed model.

5.Conclusion

In this paper, a real-time ISL hand gesture recognition with CNN and DSSD model was built

using real-time images. In the conducted experiments, we trained the proposed model for predicting 10 different words in ISL and obtained a high accuracy score even when we have used almost similar sign pose for the words 'Help' and 'Good'. This made the proposed model unique when compared with the existing methods for sign language recognition. Further research will be focused on real-time continuous sign language prediction using ISL by expanding the datasets appended with various keywords, combined with more signs from different domains.

참고문헌

[1] R. A. Kadhim and M. Khamees, "A real-time american sign language recognition system using convolutional neural network for real datasets," *TEM J.*, vol. 9, no. 3, pp. 937- 943, 2020, doi: 10.18421/TEM93-14.

[2] "https://www.ai-media.tv/sign-language-alphabets-from-around-the-world/#:~:text=There%20are%20somewhere%20between%20138,sign%20language%20as%20each%20other." .

[3] N. K. Bhagat, Y. Vishnusai, and G. N. Rathna, "Indian Sign Language Gesture Recognition using Image Processing and Deep Learning," *2019 Digit. Image Comput. Tech. Appl. DICTA 2019*, 2019, doi: 10.1109/DICTA47822.2019.8945850.

[4] N. Salem, S. Alharbi, R. Khezendar, and H. Alshami, "Real-time glove and android application for visual and audible Arabic sign language translation," *Procedia Comput. Sci.*, vol. 163, pp. 450- 459, 2019, doi: 10.1016/j.procs.2019.12.128.

[5] A. Kuznetsova, L. Leal-Taixé, and B. Rosenhahn, "Real-time sign language recognition using a consumer depth camera," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 83- 90, 2013, doi: 10.1109/ICCVW.2013.18.

[6] O. Koller, H. Ney, and R. Bowden, "Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 3793- 3802, 2016, doi: 10.1109/CVPR.2016.412.

[7] T. Raghuveera, R. Deepthi, R. Mangalashri, and R. Akshaya, "A depth-based Indian Sign Language recognition using Microsoft Kinect," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 45, no. 1, pp. 1- 13, 2020, doi: 10.1007/s12046-019-1250-6.

[8] P. S. Neethu, R. Suguna, and D. Sathish, "An efficient method for human hand gesture detection and recognition using deep learning convolutional neural networks," *Soft Comput.*, vol. 24, no. 20, pp. 15239- 15248, 2020, doi: 10.1007/s00500-020-04860-5.

[9] N. Adaloglou *et al.*, "A Comprehensive Study on Sign Language Recognition Methods," *arXiv*, pp. 1- 12, 2020.

[10] M. Hurroo and M. Elham, "Sign Language Recognition System using Convolutional Neural Network and Computer Vision," vol. 9, no. 12, pp. 59- 64, 2020.

[11] R. Elakkiya, "Machine learning based sign language recognition: a review and its research frontier," *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789, 2020, doi: 10.1007/s12652-020-02396-y.

[12] B. Olimov, S. Karshiev, E. Jang, S. Din, A. Paul, and J. Kim, "Weight initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model," in *Concurrency Computation* , 2020, doi: 10.1002/cpe.6143.