Communication through Hands in Sign Language - A CNN Collaborative Study

OPEN ACCESS

Volume: 11

Special Issue: 3

Month: July

Year: 2024

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Received: 19.05.2024

Accepted: 15.06.2024

Published: 08.07.2024

Citation:
Hemamalini, D., et al. "Communication through Hands in Sign Language - A CNN Collaborative Study." *Shanlax International Journal of Arts Science and Humanities*, vol. 11, no. S3, 2024, pp. 40–48.

DOI:
https://doi.org/10.34293/sijash.v11iS3-July.7917

Hemamalini. D
Assistant Professor, Department of AI & DS, Arjun College of Technology

Paluru Pavan Kumar Reddy
Department of AI& DS, Arjun College of Technology

Thota Nikhil
Department of AI& DS, Arjun College of Technology

Minchala Vinay Kumar
Department of AI& DS, Arjun College of Technology

Abstract
A system of communication called sign language makes use of visual motions and signals. The only form of communication for the deaf and dumb community and others with hearing impairments is sign language. Understanding sign language is so much difficult for a normal person. As a result, connecting with the wider public has always been extremely difficult for the minority community. In this study, we suggested a novel deep learning-based method for identifying sign language that can help normal and deaf individuals communicate more easily. In order to identify real-time sign language, we first created a dataset with 11 sign terms. Our bespoke CNN model was trained using these sign words. Prior to the CNN model being trained, we preprocessed the dataset. Our results show that the customized CNN model can attain the greatest accuracy of 98.6%.

Keywords: Sign Language, Deep Learning, CNN, Communication

Introduction

Sign Language (SL) is the principal technique by which deaf and dumb individuals communicate with one other and with their own community through hand and body motions. It is distinct from spoken or written language in terms of vocabulary, meaning, and grammar. In order to transmit meaningful messages, spoken language is made up of articulate sounds that are mapped onto certain words and grammatical combinations. Visual hand and body motions are used in sign language to transmit important messages. There are currently between 138 and 300 distinct varieties of sign language in use worldwide. Of the approximately 7 million deaf people in India, there are only around 250 licensed sign language interpreters. Given the shortage of sign language interpreters in the world today, it would be difficult to provide sign language to the deaf and dumb. The goal of sign language recognition is to translate these hand motions into the appropriate spoken or written language. These days, it is very common to build State of the Art (SOTA) models in Computer Vision and Deep Learning. These hand motions may be classified, and matching text can be generated, with the use of Deep Learning algorithms and Image Processing. An illustration of how the English letter "A" is used in speech or writing.

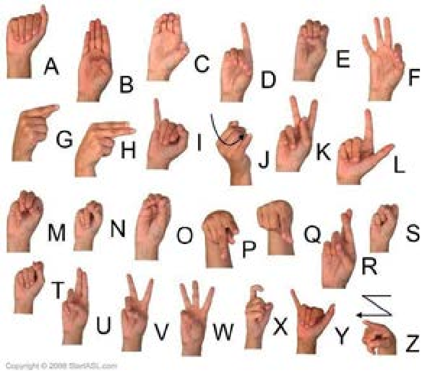


Figure 1 Sign Language Hand Gestures

In-Deep Learning The most extensively used neural network method for image and video applications is Convolution Neural Networks (CNN). We may employ sophisticated designs for Convolution Neural Networks (CNN), such as LeNET-5 and MobileNetV2, to reach the State of the Art (SOTA). All of these designs may be used, and we can utilize neural network ensemble techniques to integrate them. By doing this, we are able to create a model that can detect hand motions with about 100% accuracy. This paradigm will be implemented in standalone applications, embedded devices, or web frameworks like Django.

The live camera recognizes the hand signals, which are then translated into text. This technology will facilitate easy communication for dumb and deaf individuals.

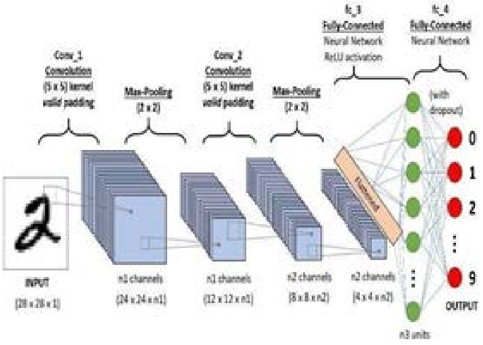


Figure 2 Convolution Neural Networks

Motivation

Developing a Sign Language Recognition system has several benefits:

- Including utilizing Sign Language hand gestures with text/speech translation or dialog systems in public spaces like airports, post offices, and hospitals.
- Sign Language Recognition (SLR) facilitates communication between normal and deaf individuals by translating videos to text or voice.

Problem Statement

Many gestures are used in sign language to give the impression that it is movement language, which is made up of various hand and arm motions. There are hand gestures and sign languages specific to each country. It should be noted that certain terms that are not well-known can be translated by merely making motions for each letter in the word. Additionally, each letter in the English vocabulary and every number from 0 to 9 has a specific gesture in sign language. These sign languages can be divided into two categories: dynamic gestures and static gestures. While the dynamic gesture is utilized for particular concepts, the static gesture is used to indicate the alphabet and numbers. Additionally, words, sentences, etc. are dynamic. Hand movements make up the static gesture, while head, hands, or both can move in the latter. Three main components make up sign language, which is a visual language: non-manual characteristics, word- level sign vocabulary, and finger spelling. While the latter is keyword-based, finger spelling is used to spell words letter by letter and convey the meaning. Though there have been several research efforts over the past few decades, designing a sign language translator remains extremely difficult. Furthermore, the look of even the same signs varies greatly depending on the signer and the perspective. This work focuses on using a Convolutional Neural Network to create a static translator from sign language. We developed a low-power network that can be utilized with web apps, standalone apps, and embedded devices that have limited resources.

Objectives

This project's primary goals are to advance the fields of automatic sign language recognition and voice or text translation. We are concentrating on hand movements in static sign language for our project. This work focused on using Deep Neural Networks (DNN) to recognize hand motions that included 10 numerals (0-9) and 26 English alphabets (A-Z). We developed a convolution neural network classifier that can identify English letters and numbers using hand gestures. The neural network has been trained using a variety of setups and designs, including LeNet-5, MobileNetV2, and our own design. To get the highest level of model accuracy, we employed the horizontal voting ensemble technique. To test our findings from a web application, we have also developed one using the Django Rest Frameworks.

Literature Review

Convolutional Neural Networks are used to Recognize Fingerspellings in Sign Language in Real Time using Depth Maps

The static fingerspelling of American Sign Language is the main topic of this study. a technique for putting into practice a system that converts sign language to text or voice without the need for sensors or portable gloves by continually recording gestures and translating them to speech.

Only a small number of photos were taken for identification using this approach. The creation of a communication tool for people with physical disabilities.

Creation of a Communication Tool for People with Disabilities

Under the MATLAB environment, the system was constructed. There are two key phases to it: the training phase and the testing phase. The author employed feed-forward neural networks throughout the training stage. The issue at hand is to MATLAB's inefficiency and the challenge of fully integrating the concurrent qualities.

Twenty of the 24 static ASL alphabets could be recognized using the processes under discussion. The occlusion issue made it impossible to distinguish the alphabets A, M, N, and S. Only a small selection of the photographs have been utilized.

Implementation

Dataset

We have used multiple datasets and trained multiple models to achieve good accuracy.

ASL Alphabet

The data is a collection of images of the alphabet from the American Sign Language, separated into 29 folders that represent the various classes.

The training dataset consists of 87000 images which are 200x200 pixels. There are 29 classes of which 26 are English alphabets A-Z and the rest 3 classes are SPACE, DELETE, and, NOTHING. These 3 classes are very important and helpful in real-time applications.

Sign Language Gesture Images Dataset

The dataset consists of 37 different hand sign gestures which include A-Z alphabet gestures, 0-9 number gestures, and also a gesture for space which means how the deaf (hard hearing) and dumb people represent space between two letters or two words while communicating.

Each gesture has 1500 images which are 50x50 pixels, so altogether there are 37 gestures which means there 55,500 images for all gestures. Convolutional Neural Network (CNN) is well suited for this dataset for model training purposes and gesture prediction.

Data Pre-Processing

An image is nothing more than a 2-dimensional array of numbers or pixels which are ranging from 0 to 255. Typically, 0 means black, and 255 means white. Image is defined by mathematical function f(x, y) where „x‟ represents horizontal and „y‟ represents vertical in a coordinate plane. The value of f(x, y) at any point is giving the pixel value at that point of an image.

Image Pre-processing is the use of algorithms to perform operations on images. It is important to Pre- process the images before sending the images for model training. For example, all the images should have the same size of 200x200 pixels. If not, the model cannot be trained.

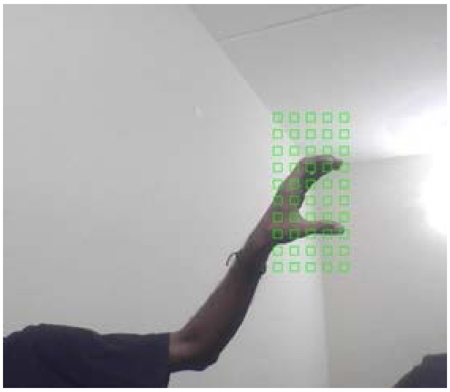


Figure 3 Sample Image without Pre-Processing

The steps we have taken for image Pre-processing are:

- Read Images.
- Resize or reshape all the images to the same
- Remove noise.
- All the image pixels arrays are converted to 0 to 255 by dividing the image array by 255.

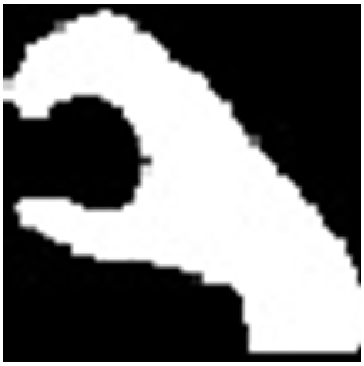


Figure 4 Pre-Processed Image

Convolution Neural Networks (CNN)

Computer Vision is a field of Artificial Intelligence that focuses on problems related to images and videos. CNN combined with Computer vision is capable of performing complex problems.

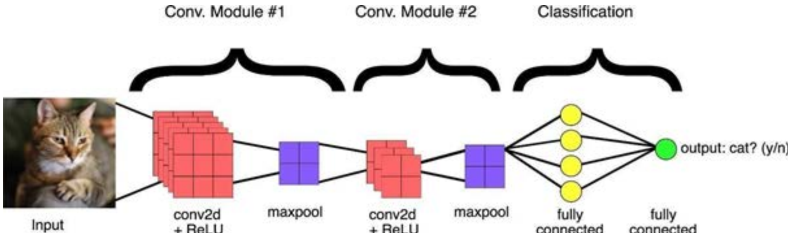


Figure 5 Working of CNN

The Convolution Neural Networks has two main phases namely feature extraction and classification. A series of convolution and pooling operations are performed to extract the features of the image.

A fully connected layer in the convolution neural networks will serve as a classifier. In the last layer, the probability of the class will be predicted. The main steps involved in convolution neural networks are:

1. Convolution
2. Pooling
3. Flatten
4. Full connection

Convolution

Convolution is nothing but a filter applied to an image to extract the features from it. We will use different filters to extract features like edges, highlighted patterns in an image. The filters will be randomly generated.

What this convolution does is, creates a filter of some size says 3x3 which is the default size. After creating the filter, it starts performing the element- wise multiplication starting from the top left corner of the image to the bottom right of the image. The obtained results will be extracted feature.

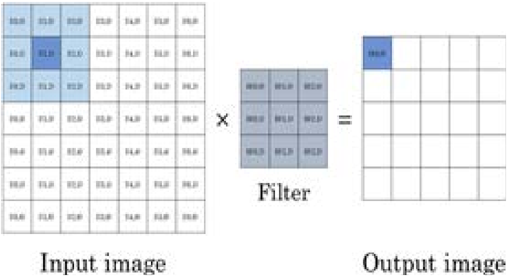


Figure 6 Convolution

The size of the output matrix decreases as we keep on applying the filters.
Size of new matrix = (Size of old matrix — filter size) +1.

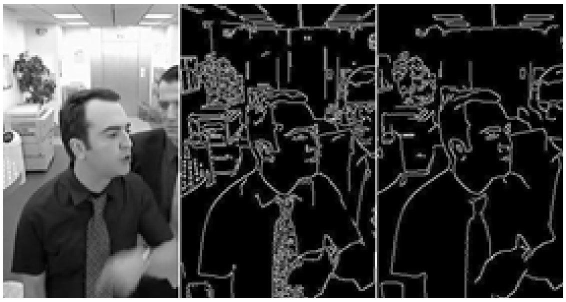


Figure 7 Feature Extraction

Pooling

After the convolution operation, the pooling layer will be applied. The pooling layer is used to reduce the size of the image. There are two types of pooling:

1. Max Pooling
2. Average Pooling

Max pooling

Max pooling is nothing but selecting the maximum pixel value from the matrix.

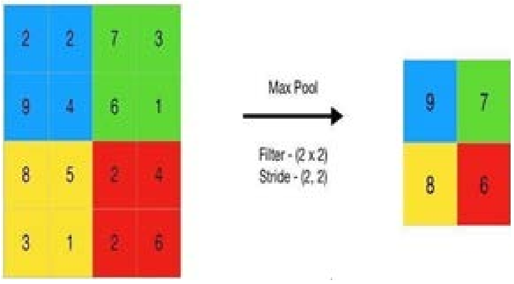


Figure 8 Max Pooling

This method is helpful to extract the features with high importance or which are highlighted in the image.

Average Pooling

Unlike Max pooling, the average pooling will take average values of the pixel.

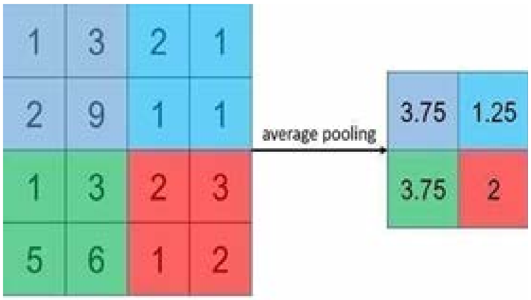


Figure 9 Average Pooling

In most cases, max pooling is used because its performance is much better than average pooling.

Flatten

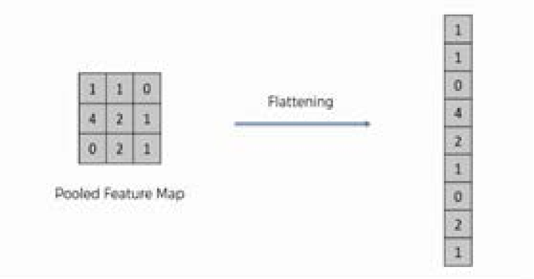


Figure 10 Flatten

The obtained resultant matrix will be in muti- dimension. Flattening is converting the data into a 1- dimensional array for inputting the layer to the next layer. We flatten the convolution layers to create a single feature vector.

Full Connection

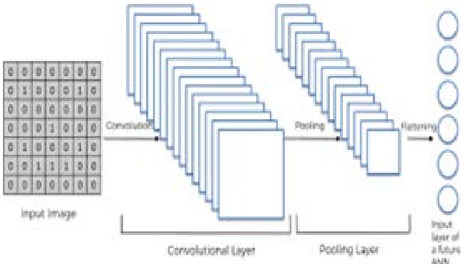


Figure 11 Full Connection

A fully connected layer is simply a feed-forward neural network. All the operations will be performed and prediction is obtained. Based on the ground truth the loss will be calculated and weights are updated using gradient descent backpropagation algorithm.

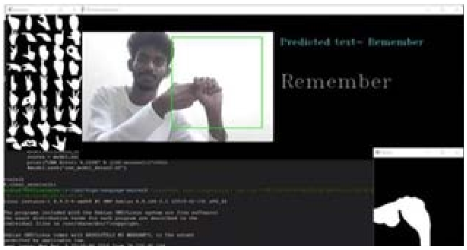
Experimental Results

We have trained all the models for around 10-15 epochs with a batch size of 32.

All the models performed well on the test cases. After applying the horizontal voting ensemble technique for these 3 models, we have achieved almost 100% accuracy.

We have used OpenCV to test our results in the live camera. This is a sample result on a live camera.

Sample Output for the word "Remember"



Predicted text- Remember

Remember

Conclusion

In conclusion, we were successfully able to develop a practical and meaningful system that can able to understand sign language and translate that to the corresponding text. There are still many shortages of our system like this system can detect 0-9 digits and A-Z alphabets hand gestures but doesn‟t cover body gestures and other dynamic gestures. We are sure and it can be improved and optimized in the future.

References

1. Brill R. 1986. The Conference of Educational Administrators Serving the Deaf: A History. Washington, DC: Gallaudet University Press.
2. Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.
3. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510-4520, doi:10.1109/CVPR.2018.00474.
4. L. K. Hansen and P. Salamon, “Neural network ensembles,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 10, pp. 993-1001, Oct. 1990, doi: 10. 1109/34.58871.
5. Kang, Byeongkeun, Subarna Tripathi, and Truong Q. Nguyen. “Real- time sign language fingerspelling recognition using convolutional neural networks from depth map.” arXiv preprint arXiv: 1509.03001 (2015).
6. Suganya, R., and T. Meeradevi. “Design of a communication aid for phys- ically challenged.” In Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, pp. 818- 822.IEEE, 2015.

7. Sruthi Upendran, Thamizharasi. A, “American Sign Language Interpreter System for Deaf and Dumb Individuals”, 2014 International Conference on Control, Instrumentation, Communication and Computa.
8. David H. Wolpert, Stacked generalization, Neural Networks, Volume 5, Issue 2, 1992, Pages 241-259, ISSN 0893-6080, https://doi.org/10.1016/S0893- 6080(05)80023-1.
9. Y. Liu, X. Yao, Ensemble learning via negative correlation, Neural Networks, Volume 12, Issue 10,1999, Pages 1399-1404, ISSN 0893-6080, https://doi.org/10.1016/S0893- 6080(99)000 73-8.
10. MacKay D.J.C. (1995) Developments in Probabilistic Modelling with Neural networks — Ensemble Learning. In: Kappen B., Gielen S. (eds) Neural Networks: Artificial Intelligence and Industrial Applications. Springer, London. https://doi.org/10.1007/978-1-4471-3087- 1_37.