Real-time American Sign Language Gesture Recognition using Convolutional Neural Networks (CNNs)

Ronil Christian
Illinois Institute of Technology
Chicago, IL, USA
rchristian@hawk.iit.edu

Abstract

American Sign Language has become a complex and important means for communication for the deaf and dumb community. With the increasing popularity of technology, the need has arisen for the development of automatic recognition of ASL gestures. Hereby presented is a deep learning approach for the classification of ASL gestures using convolution neural networks (CNNs). Overall, this paper aims to show promise for automatic recognition of ASL gestures which is fully generalizable over a wide range of applications.

Keywords: *Convolutional Neural Networks, Deep Learning*

1. Introduction

American Sign Language (ASL) is a visual language that is used by the deaf and hard of hearing community in the United States and Canada. According to the National Institute on Deafness and Other Communication Disorders, approximately 15% of American adults (37.5 million people) report some trouble hearing, and around 2 to 3 out of every 1,000 children in the United States are born with a detectable level of hearing loss.

ASL gesture classification remains a challenging problem due to the complexity of ASL gestures and the high degree of variation in signing styles among individuals. Automatic recognition of ASL gestures has the potential to provide a more efficient and accurate means of communication for the deaf and hard of hearing community.

This paper intends to present an ASL recognition system that uses Convolutional Neural Networks (CNN) in real time to translate a video of a user's ASL signs into text. Our problem consists of three tasks to be done in real time:

1. Obtaining video of the user signing (input)
2. Classifying each frame in the video to a letter
3. Reconstructing and displaying the most likely word from classification scores (output)

From a computer vision perspective, this problem represents a significant challenge due to a number of considerations, including:

- Environmental concerns (e.g., lighting sensitivity, background, and camera position)
- Occlusion (e.g., some or all fingers, or an entire hand can be out of the field of view)
- Sign boundary detection (when a sign ends and the next begins)
- Co-articulation (when a sign is affected by the preceding or succeeding sign)

While Neural Networks have been applied to ASL letter recognition in the past with accuracies that are consistently over 90%, many of them require a 3-D capture element with motion-tracking gloves or a Microsoft Kinect, and only one of them provides real-time classifications. The constraints imposed by the extra requirements reduce the scalability and feasibility of these solutions.

Our system features a pipeline that takes video of a user signing a word as input through a web application. We then extract individual frames of the video and generate letter probabilities for each using a CNN (letters a through y, excluding j and z since they require movement). With the use of a variety of heuristics, we group the frames based on the character index that each frame is suspected to correspond to. Finally, we use a language model in order to output a likely word to the user.

2. Related Work

Researchers have used a variety of classifiers like liner classifiers, neural networks, and Bayesian networks.

Recent studies have shown that deep learning approaches, particularly CNNs, can achieve state-of-the-art performance in gesture recognition tasks. For example, in a study by Donahue et al. (2015), a deep learning approach was used to classify ASL gestures with an accuracy of 90.1%. Similarly, in a study by Zhang et al. (2017), a CNN-based approach achieved an accuracy of

96.7% on a dataset of Chinese Sign Language (CSL) gestures.

Other studies have focused on improving the robustness of gesture recognition systems by incorporating multiple modalities, such as depth information and skeletal tracking. For example, in a study by Hu et al. (2018), a multimodal approach was proposed that combined RGB images and depth maps for recognition of ASL gestures.

Mekala et al. classified video of ASL letters into text using advanced feature extraction and a 3-layer Neural Network. They extracted features in two categories: hand position and movement. Prior to ASL classification, they identify the presence and location of 6 "points of interest" in the hand: each of the fingertips and the center of the palm.

3. Approach and Methods

3.1. Classifier Development

Since Convolutional Neural Networks (CNNs) have seen incredible success in handling tasks related to processing images and videos, this paper adopts the same machine learning algorithm to classify the ASL gestures.

A primary advantage of utilizing such techniques stems from CNNs abilities to learn features as well as the weights corresponding to each feature. Like other machine learning algorithms, CNNs seek to optimize some objective function, specifically the loss function.

Using a softmax-based classification head allows us to output values akin to probabilities for each ASL letter. These probabilities afforded to us by the softmax loss allow us to more intuitively interpret our results and prove useful when running our classifications through a language model.

3.2. General Technique

* Code file attached with this report representing a baseline model that has been created. Further tuning required.

3.3. Developing the pipeline

4. Dataset Features

4.1. Dataset Description

The dataset format is patterned to match closely with the classic MNIST. Each training and test case represents a label (0-25) as a one-to-one map for each alphabetic letter A-Z (and no cases for 9=J or 25=Z because of gesture motions). The training data (27,455 cases) and test data (7172 cases) are approximately half the size of the standard MNIST but otherwise similar with a header row of label, pixel1, pixel2, …, pixel784 which represent a single 28x28 pixel image with grayscale values between 0-255.

4.2. Data Pre-processing

* Yet to be done

5. Experiments, Results, and Analysis

5.1. Evaluation Metrics

* Yet to be done

5.2. Experiments

* Yet to be done

5.3. Results

* Yet to be done

5.4. Discussion

* Yet to be done

5.4.1 Loss and Accuracy

5.4.2 Confusion Matrices

5.4.3 Real-time user testing

6. Conclusions and Future work

6.1. Conclusion

* Yet to be done

6.2. Future Work

* Yet to be done

References

[1] Garcia B., Viesca S.A. Real-time American sign language recognition with convolutional neural networks. Convolutional Neural Netw. Vis. Recognit. 2016;2:225–232. [Google Scholar]

[2] Abu-Jamie, Tanseem N., and Samy S. Abu-Naser. "Classification of Sign-Language Using Deep Learning by ResNet." (2022).

[3] P. Mekala et al. Real-time Sign Language Recognition based on Neural Network Architecture. System Theory (SSST), 2011 IEEE 43rd Southeastern Symposium 14-16 March 2011.

[4] https://www.kaggle.com/datasets/datamunge/sign-language-mnist

[5] https://www.tensorflow.org/api_docs

[6] https://docs.opencv.org/4.x/

NOTE: This is document is not complete. This is some work done until now (along with the code files attached). The topics represent an outline of the work that will done for completion.