Sign Language Recognition Using Deep Learning

Author
MHD Khaled Maen
mk.maen93@gmail.com

Supervisor
Assoc. Prof Dr. Amelia Ritahani Ismail
amelia@iium.edu.my

Abstract

Communication is an essential part of our life. Unfortunately, some of us were born in various types of disability such as deaf, since hearing impaired people cannot listen they loosed the ability to learn how to speak so they developed a new communication way to interact with other people by using distinct hand gestures, which was not enough to overcome this issue since most of hearing people do not understand sign language, even now with all technologies and tools it is remain a challenging problem to solve. For the mentioned reason, the intention of this paper is to improve the ordinary model to translate the sign language gestures into a voice. In that, Deep learning is remarkably serviceable for this mission, first by detecting a hand in a video frame using Convolutional Neural Network (CNN) algorithm followed by recognizing the letter and state the matching sound. The accuracy achieved for a hand gesture detection using CNN model is more than 90% for all the proposed signs.

1. Introduction

Communication is a process of sending and receiving data among individuals. People communicate with each other's by too many ways yet the best one is by talking. Numerous researchers trusted that the significance of communication is like the importance of breathing. Indeed, communication facilitates the spread of knowledge and structures connections between individuals.

Deep learning added an immense lift to the already rapidly developing field of computer vision. Moreover, a lot of new utilization of computer vision techniques have been presented and they are currently ending up some portion of our regular day to day existence.

Alongside with the intensity of the present computers, there are now various algorithms that were developed to empower computers to perform tasks such as object tracking and pattern recognition.

In this paper, the attention will be on hand gestures detection and make an interpretation of them into voice.

2. Related Work

In spoken languages words being pronounced by vocal cords which produce sound wave can be heard, however for deaf people this sounds are not heard [1], so sign language is based on the visual tools to communicate and receive information.

American Sign Language (ASL) is completely different from English. It has the basic features of a language such as the letters as shown in Figure 1.

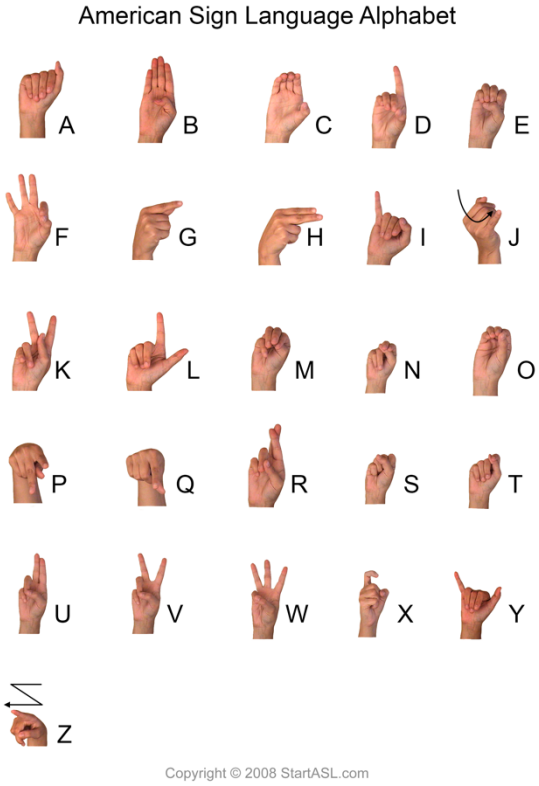

Figure 1- American Sign Language

Following the same path many previous attempts has been conducted to find a solution for this problem, marvelous work has been done, but most of them was just a prove of concept were the authors concentrate on a few signs to be detected.

Starting with Bao, Maqueda, del Blanco, & Garc´ıa They proposed a Deep convolutional neural network algorithm for hand-gesture recognition for 7 different gestures without hand localisation, since the hands only occupy about 10% of the image. They used a combination of 9 convolution layers, 3 fully connected layers, interlaced with Rectified Linear Unit (ReLU) and dropout layers. Alongside this architecture the apply some image processing techniques to have sufficient computation efficiency and memory requirement. According to the paper the accuracy achieved was 97.1% in the images with simple backgrounds and 85.3% in the images with complex backgrounds [2].

The second work conducted by Pyo, Ji, You, & Kuc they introduced a depth-based hand data with convolution neural networks (CNNs).The hand gesture dataset has roughly 6,000 RGB-D images in each of 12 labels. In all, there are approximately 60,000 training images, 1S,000 validation images, and 12,000 training images [3].

One more work was done by Devineau, Moutarde, Xi, & Yang they introduced a 3D hand gesture recognition approach based on a deep learning model using Convolutional Neural Network (CNN). The proposed model only uses hand-skeletal data and no depth image. The model produced by multi-channel convolutional neural network with two feature extraction modules and a residual branch per channel 2.6. The achieved accuracy was a 91.28% classification accuracy for the 14 gesture classes case and an 84.35% classification accuracy for the 28 gesture classes case [4].

.

3. Methodology

In order to get a model with satisfying results, we setup an environment with a powerful computational resources and large datasets.

Figure 2 - System flowchart

3.1. Environment

Google Cloud Platform **(**GCP**)** is a cloud computing service provided by Google which provides a lot of products including cloud computing.

Google Compute Engine (GCE) is the Infrastructure as a Service (IaaS) enables users to setup Virtual Machines (VMs) can be accessed via the developer console, command line interface (CLI) and RESTful API [5].

The VM instance of choice run on an Ubuntu operating system version 16.04. Alongside, the VM instance is powered by 8 Graphics Processing Units (GPUs) of type NVIDIA Tesla K80 which engineered to boost throughput in real-world applications by 5-10x times, compared to Central Processing Unit (CPU).

3.2. Data set

The data set is a collection of images for American Sign Language (ASL) alphabet, contained of 87,000 images of 200x200 pixels frame size as shown in Figure 3. There are 29 classes, the first 26 classes are the letters A to Z and the last three classes are space, delete and empty signs [6].

Figure 3 – The letter K in American Sign language

3.3. Image processing

In order to extract useful information from an image, some methods need to be implemented:

3.3.1. Image annotation

The process of annotating and labeling every object in the image in this case the sign has been labeled, so the machine can identify the important part of the image which has the object and the characteristics of it.

The data of annotated image saved in eXtensible Mark-up Language (XML) file such as object class and the bounding box coordination.

3.3.2. XML to Comma Separated Values (CSV)

The goal of this process is to convert the semi-structured data stored in the XML files into structured data stored in one file CSV so it can be transferred into TensorFlow(TF)Records.

3.3.3 CSV to TFRecords

Since the library used to implement the algorithm for the experiment is TensorFlow it requires a certain type of data, so it can read efficiently from it since the data being serialized and stored in a sequence of binary records.

3.4 TensorFlow Object Detection API

Open source framework built on top of TensorFlow in order to help researchers to construct, train and deploy object detection models [7].

3.5 Algorithm

The algorithm used in this experiment is MobileNet-SSD as shown in Figure 4 [8]; MobileNet is a base network which use depth-wise separable convolutions to build light weight deep neural networks provide high level features for classification or detection [9], after removing the fully connected layer at the end of the network and replace it with Single Shot Multibox Detector (SSD) one of the most popular algorithm in object detections [10]. SSD generates a box with score of presence of each object category and adjusts the box to better match the object shape. In addition SSD eliminates proposal generation and feature resampling stage by encapsulates all computation in a single network, alongside with the ability to combines predictions from multiple features maps with different resolutions to naturally handle objects of various sizes.

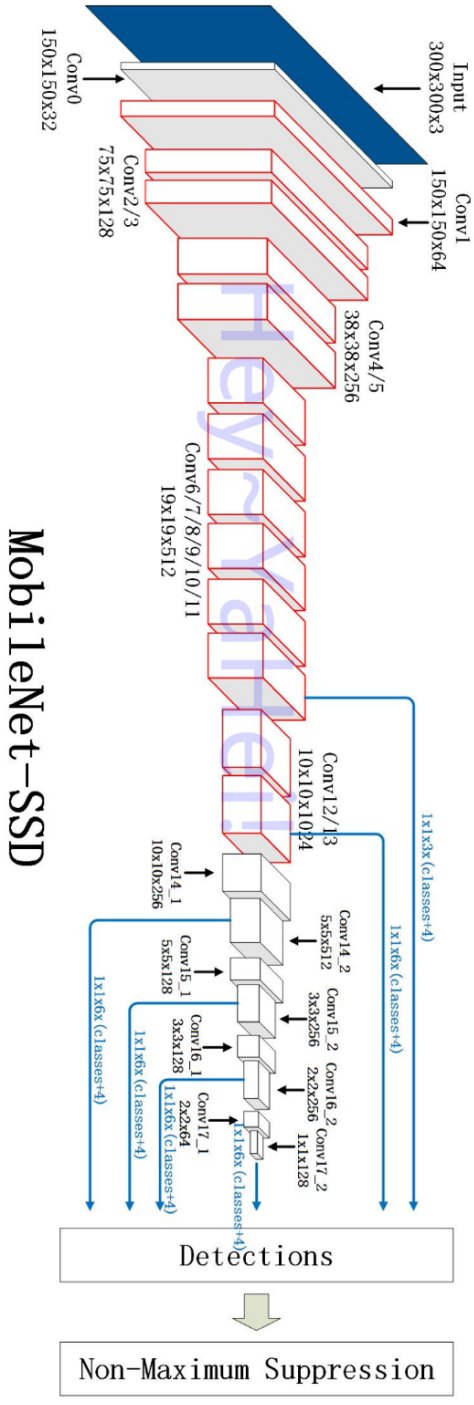

Figure 4 - MobileNet-SSD architecture

4. Results

After 13587 steps the model being able to detect and identify each and every sign of the 29 signs including the empty one with classification loss 0.143 and localization loss 0.0327.

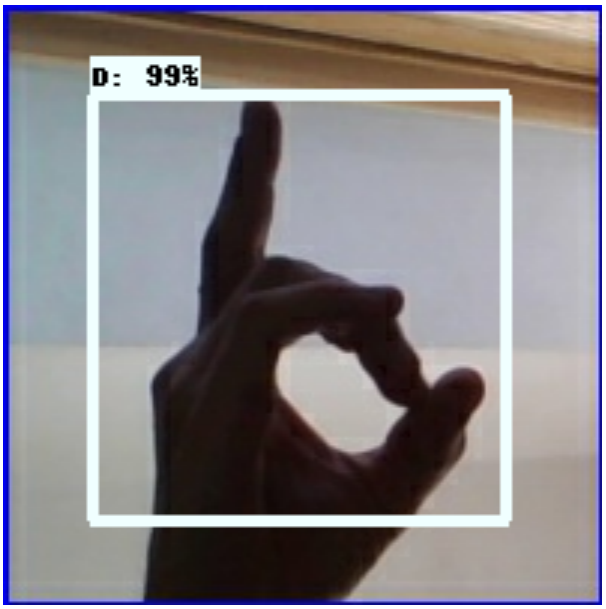

Figure 5 - Prediction of letter D

5. Feature works

One of the potential improvements is to collect data for complete words and build a model which translate the voice into signs to make it two-ways connections.

6. References

[1] K. Ishizaka and J. L. Flanagan, "Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords," *Bell Syst. Tech. J.*, 1972.

[2] P. Bao, A. I. Maqueda, C. R. Del-Blanco, and N. Garciá, "Tiny hand gesture recognition without localization via a deep convolutional network," *IEEE Trans. Consum. Electron.*, vol. 63, no. 3, pp. 251–257, 2017.

[3] J. Pyo, S. Ji, S. You, and T. Kuc, "Depth-based hand gesture recognition using convolutional neural networks," in *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence, URAI 2016*, 2016.

[4] G. Devineau, F. Moutarde, W. Xi, and J. Yang, "Deep learning for hand gesture recognition on skeletal data," in *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 2018.

[5] "Compute Engine - IaaS | Compute Engine | Google Cloud." [Online]. Available: https://cloud.google.com/compute/.

[6] "Kaggle | ASL-Alphabet." [Online]. Available: https://www.kaggle.com/grassknoted/asl-alphabet.

[7] "TensorFlow | Object_detection." [Online]. Available: https://github.com/tensorflow/models/tree/master/research/object_detection.

[8] "MobileNet-SSD | Architecture." [Online]. Available: https://hey-yahei.cn/2018/08/08/MobileNets-SSD/index.html.

[9] A. G. Howard *et al.*, "MobileNets," *arXiv Prepr. arXiv1704.04861*, 2017.

[10] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.