Study of Convolutional Neural Network in Recognizing Static American Sign Language

Lee Yi Bin
Department of Mechanical Engineering
Faculty of Engineering and Technology
Tunku Abdul Rahman University College
Kuala Lumpur, Malaysia
yibin5243@gmail.com

Goh Yeh Huann
Department of Mechanical Engineering
Faculty of Engineering and Technology
Tunku Abdul Rahman University College
Kuala Lumpur, Malaysia
gooyh@tarc.edu.my

Lum Kin Yun*
Department of Mechanical Engineering
Faculty of Engineering and Technology
Tunku Abdul Rahman University College
Kuala Lumpur, Malaysia
lumky@tarc.edu.my

Abstract—Sign language is a form of communication language to connect a deaf-mute person to the world. It involves the use of hand gestures and body movement in order to express an idea. Nevertheless, publics are mostly not educated to comprehend the sign language. For this reason, there is a need to have a translator to facilitate the communication. This paper would like to present a Convolutional Neural Network (CNN) model for predicting American Sign Language. There are 4800 images were captured to train and validate the proposed model. 95% recognition accuracy was attained in experiment, which shows robust performance in recognition 24 static American Sign Language pattern. The successful development of this model can be served as the basis to develop a more complicated sign language translator.

Index Terms—Convolutional Neural Network, American Sign Language, Machine Learning, Computer Vision

I. INTRODUCTION

According to the statistics from the World Health Organization (WHO), there are more than 5% or 466 million people of the world’s populations are suffering from hearing loss. Hearing loss or hearing impairment, can be either partial or total inability to hear. If this happens during childhood, it can hinder their language learning ability and results in language impairment or so called hearing mutism. A person with hearing mutism may have a distorted speech or may not speak at all [1]. This creates a communication barrier between them and the society.

This special group of people is not being left out with the aid of sign language. Sign language is a kind of visual communication language which uses a combination of the hand gesture, facial expression and body posture. It solves the general communication problem within their community. But there is no one size fits all solution in sign language. Different sign languages are employed in different countries or regions. For instance, Americal Sign Language (ASL) is distinct from British Sign Language (BSL). One particular sign language literacy may not be able to communicate with another person who learned different type of sign language.

In addition, most general publics are illiterate in any form of sign language. Unless there is a deaf-mute person surrounds them, most tend to not interested and ignore the importance of sign language. Even though some devices such as a data glove [2] and wearable surface EMG [3], [4] were built to serve this communication purpose, they can be costly, cumbersome and inconvenient when come to practical. For example, sign language performer needs to wear this special glove which used to detect the hand’s motion and posture then translate it into words or voice.

It is important to continuously work out for this group of people who require our special attention so that they can live normally like you and me. There are many researches being conducted to find out a cheaper and convenient alternative [5]. Since hand sign gestures are perceived through vision, computer vision can be a subject of profound interest. With the widely available smartphone with camera, one of the solutions is to develop a vision based sign language recognition system. Traditionally, vision based systems are generally subject to reliability issues. Background noise, colours, and lighting are greatly vary under real environment. This would result a lower detection rate. However, the recent breakthrough in computer vision and machine learning technology has shown a huge progress in image classification [6], [7]. Conventional machine learning relies on manual features extraction. Instead of constructing complex handcrafted features, CNNs are able to automate the process of feature construction. This would greatly reduce the human error and improve the detection accuracy.

This paper will put forward a study on the application of CNN in recognizing the American Sign Language. A CNN model is built to recognize the American Sign Language images captured by a smartphone camera. The scope is limited to static alphabet recognition at this stage. Excluding two characters which involve hand’s movement, a total of 24 characters were chosen. A detection accuracy of 95% was recorded.

The rest of this paper is organized as follows: Some related work will be discussed in Section II. Then the methodology and details of Convolutional Neural Network are presented in Section III. Experimental results are discussed in Section IV. Finally, the finding of this paper is concluded in Section V with some recommendations for future work.
II. RELATED WORK

Sawant Pramada et. al. from India had been focusing on image processing and template matching technique to perform recognition on the Binary Sign Language. It is a form of sign language represents the alphabetical characters in the form of binary. For instance, A is represented by \(00001_2\), B is \(00010_2\) and so on. All five binary numbers is indicated with one hand’s fingers. The image with neutral background is captured by a webcam and converted into grayscale image. Then, it will be compared with the stored coordinate map template and display the matched pattern in the textual or audio form. Due to the simplicity of Binary Sign Language, a high recognition accuracy of 90.9\% was achieved [8].

Vivek Bheda and N. Dianna Radpour suggested a Deep Convolutional Networks classification algorithm of American Sign Language for alphabets and digits. Common CNN architecture was used which consists of 3 groups of 2 convolutional layers followed by a max-pool layer and a dropout layer, as well as two groups of fully connected layer followed by a dropout layer and an output layer. They trained this system from scratch and produce 82.5\% accuracy on validation sign language recognition system [9].

Lionel Pigou et. al. proposed a predictive model to recognize 20 Italian Sign Language performed by 27 users with variations in surroundings, clothing, lighting and gesture movement recorded with a Microsoft Kinect. CNN automated feature construction was proposed as opposed to the traditional handcrafted feature extraction. Two CNNs were used to extract hand features and upper body features, respectively. An accuracy of 91.7\% was recorded [10].

Garcia and Viesca proposed the idea of translating the American Sign Language (ASL) into alphabet using CNN with transfer learning. They use colour images of America Sign Language as training set for CNN. A pre-trained GoogLeNet architecture was used in this design. Due to the similarity in some signs, it was only around 70\% accuracy was recorded for all 24 letters. Then, they decreased the number of classes and attained a validation accuracy of nearly 98\% with five letters and 74\% with ten [11].

For motion based sign language recognition system, the challenge lies in continuous tracking of hand regions, segmenting hand-shape images from the background and moving gesture recognition. Variation and occlusion of hands and body joints will increase the design complexity to obtain a satisfying result. Jie Huang, et.al. proposed a three input capturing system consists of color images, depth images and body skeleton images captured by Microsoft Kinect. Instead of the traditional hand-crafted features extraction, he introduced 3D convolutional neural network (CNN) for an automatic features extraction. The highest recorded accuracy for his work is up to 94.2\% [12]. However, motion based sign language is beyond the scope of discussion of this paper. Moreover, Microsoft Kinect is not available everywhere. Thus, we will be only focusing on static sign language recognition which can be easily done by using smartphone camera.

III. METHODOLOGY

A. Dataset

Preparing enough dataset is a very prominent part in machine learning. This work would only focus on the static Americal Sign Language gestures which are letters A to Y as showed in Fig. 1. Letters J and Z were excluded since they require movement. There were 24 gestures images were collected using a smartphone camera to form the dataset. 200 images were captured for each gesture from two different users under the variation of background and lighting condition. There was a total of 4800 images. To reduce the training time, all the images were resized to 32 x 32. Some examples of the captured images were displayed in Fig. 2.

![Fig. 1. American Sign Language for all alphabets.](image)

![Fig. 2. Some example of captured and resized dataset. From left to right (top) A, B, C and (bottom) D, E, F.](image)

There were 180 images used for training and the remaining 20 were used for testing. In the other words, 90\% of the dataset were used as training set and 10\% of the dataset were used to validate the accuracy of the sign language recognition system.

B. Convolutional Neural Network (CNN)

There are four major components needed to form a CNN model excluding the input and output layer. They known as
convolution layers, pooling/subsampling layers, flattening, and fully connected layers.

The convolutional layers serve as feature extractors to learn the feature representations of their input images. It is otherwise known as filters to detect image features such as lines, edges, colors and other visual elements. There are convolutional neurons to perform convolutional operations by scanning over every pixels of the input image and make way for the result to the next layer. A convolutional layer can have many filters to detect different features. All the filters' weights will be updated in backpropagation. A complete CNN can have more than one convolutional layer to further process the extracted features to form more complex features.

The next layer in the convolutional network is generally known as max pooling layer. The main function of max pooling is to further reduce the size of images. For example, max pooling takes the largest value from one patch of an image, after that places the largest number into a new matrix, so that it discards the rest of the information that contains in activation map.

The next layer in a convolutional neural network is the flattening layer. The function of this layer is to convert all the pooled images into a continuous vector through flattening. For example, it will convert all the two dimensional arrays into a single long continuous linear vector to serve as the inputs of the next fully connected layer.

Lastly, the fully connected layer is the neural network to perform classification based on the extracted inputs from the convolution layers.

C. Proposed Architecture

A CNN model is built which consists of one input layer, 2 cascaded convolutional layers with max pooling and dropout, one flattening layer, one fully connected layer with dropout, and one output layer with softmax function. These whole model is implemented in multiple layers as showed in Fig. 3. Layer 1 is having a two cascaded convolutional layers followed by a max pooling layer. Each convolutional layer is having 32 filters with a window size of 3x3. It is accompanied by a 2x2 max pooling layer. Then, layer 2 is another two convolutional layers with 64 filters, each having a window size of 3x3. Again, it is followed by a 2x2 max pooling layer. The activation function of all the convolutional layers are “ReLU” and all the max pooling layers are having a stride of 2. To reduce the training time, batch normalization was applied on every convolutional layer to normalize the activations of the previous layer. A 25% dropout was set to every layer to prevent overfitting.

Then, the flattening layer is added to convert the 3D array into a single array. After that, fully connected layer is added followed by the output layer with softmax function. Fully connected layer is the neural network to perform the classification founded on the extracted features from the previous convolutional layers. Similar batch normalization and dropout were also applied to this layer.

Finally, an output layer with softmax function is added to classify 24 classes of American Sign Languages.

D. Generalization and Training

Experiments were all conducted on a laptop with Intel Core i5-7300HQ, 8GB SDRAM and a NVIDIA GeForce GTX 1050. The whole models were implemented using the Python programming. Dropout layer and ImageDataGenerator for data augmentation were utilized to avoid overfitting.

A dropout ratio of 0.25 was set. It means that one in 4 inputs will be randomly excluded from each update cycle. Dropout is the technique randomly selected neurons and setting a fraction rate to ignore during the training. It means that one of the neurons is temporally removed on the forward pass. Any updated weight are not applied to that neuron on the backward pass. Therefore, the effect is that the network becomes less
sensitive to the specific weight of the neurons and less likely to overfit the training data.

The data augmentation was performed in real time on the CPU during the training phase while the model was being trained. The ImageDataGenerator for data augmentation consists of random zoom into images, randomly rotate the images in the range up to 0° - 180°, randomly shift the images horizontally according to the fraction of total width and randomly shift the images vertically according to the fraction of total height.

IV. EXPERIMENTAL RESULTS

The training was performed with Adam optimizer with categorical cross entropy loss function. Epoch is set to be 100 and learning rate was set at 0.003. The training loss and accuracy plot are shown in Fig. 4. The figure shows low loss with limited overfitting after 100 epochs.

![Training Loss and Accuracy](image)

Fig. 4. Training and validation loss/accuracy plot for American Sign Language CNN classifier trained with Keras.

According to the Fig. 5, we can observe that overall result is quite accurate except for the alphabet W and X. There is only 60% accuracy on alphabet W and X. This is due to the sign similarity and background noise. The overall accuracy of the test set is 95% which is the best model obtained and we observed only 5% false positive rate.

![Confusion Matrix](image)

Fig. 5. Confusion Matrix for American Sign Language every class of alphabet.

Table I shows the comparison with some previous work. By far, it is the method with the highest recorded accuracy on American Sign Language.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Description</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sawant Pramada et.al. 2013</td>
<td>Image processing and template matching technique on Binary Sign Language</td>
<td>90.9%</td>
</tr>
<tr>
<td>Vivek Bheda and N. Dianna Radpour 2017</td>
<td>Deep CNN on alphabets and digits American Sign Language</td>
<td>82.5%</td>
</tr>
<tr>
<td>Lionel Pigou et. al. 2014</td>
<td>Two CNNs on 20 Italian Sign Language to extract hand features and upper body features using Microsoft Kinect</td>
<td>91.7%</td>
</tr>
<tr>
<td>Garcia and Viesca 2016</td>
<td>CNN on 24 American Sign Language with GoogLeNet transfer learning</td>
<td>70%</td>
</tr>
<tr>
<td>Jie Huang, et.al. 2015</td>
<td>3D CNN on 25 sign language vocabularies using Microsoft Kinect</td>
<td>94.2%</td>
</tr>
<tr>
<td>This work</td>
<td>CNN on 24 American Sign Language using phone camera</td>
<td>95%</td>
</tr>
</tbody>
</table>

V. CONCLUSION AND FUTURE WORK

In this paper, we described a CNN architecture to recognize 24 letters in American Sign Language. The experimental results show the proposed method is effective in predicting static alphabetical gestures, which in return can be served as a beginning step to bridge the communication gap between the deaf-mute person and the community. As compared to most previous works which were using Microsoft Kinect, this work was merely using the dataset captured by widely available smartphone camera which gives more flexibility and convenience.

This work can be continued with real-time video-based sign language recognition to give more usability. At the same time,
this can also accommodate more sophisticated sign language which involves hand movements. To minimize the environmental noise interference, video processing and recognition involves region of interest segmentation and hand tracking will be the next research. Besides, image occlusion is not been studied in this work due to limitation of self-created database. This can be a challenging issue when part of the performing signs is not in view.

REFERENCES


