Sign of times: Classifying Sign Language with MediaPipe

Emmanuel Z. Pedernal
College of Computer Studies
emmanuel_pedernal@dlsu.edu.ph

***Abstract*—Recent events seem to bring about attention on challenged individuals, current initiatives that supports the marginalized groups includes integrating sign language to programs, and integration of hearing with non-hearing individuals, while these support are present there is still discrepancy persists between able and disabled individuals in terms of communication and opportunity. This research aims to find suitable machine learning model with the use of media pipe to use for sign language detection.**

I. Introduction

Despite advances in technology, sign language still plays pivotal role in communication for able and disabled person. Sign language is a type of visual-gestural language that came from hand movements, various facial expressions, and postures to portray messages. In this research the focus would be highlighting the capabilities of machine learning models for sign language recognition. The researcher aims to examine an alternative python library MediaPipe. MediaPipe provides real-time multimedia processing task including hand tracking and recognition of gestures. Through MediaPipe the researcher seek to create and advance methods in the field of sign language recognition.

Leveraging MediaPipe hand landmarks with parameter that focuses on a hand with most prominent gesture, a single complexity and minimum detection of 70%, the research has two keys approach; detection of hand gestures and classification of hand gestures. This research aims to develop an model for accurately detecting and tracking hands in images, the library's hand tracking module provides robust hand localization that precisely identify the region of area where the hand gestures are. Upon hand detection our goal is for the model to classify to which class the gesture is from. The research utilized Convolutional Neural network and Neural Networks to detect the class for each gesture.

II. Related Works

The paper Sign Language Recognition Using Convolutional Neural Networks [1] was trained with CLAP14 dataset, the paper describe the dataset that contains different Italian gestures, these gestures are recorded with Microsoft Kinect device. The CNN architecture consist of two CNNs for hand gesture and upper portion of the body feature extraction, followed by classical ANN for the actual classification of hand gesture. Data Augmentation and dropout was utilized prior to the training of data this resulted in the model achieving high accuracy rates on validation and test sets with results of error rate 18.9% prior to using ReLU and LCN down to 8.30% an improvement of almost 20%. The strategic design of the architecture by incorporating CNN's to divide the critical feature extraction the model can exploit the hierarchical relationships within the data which creates robust feature learning in addition the research used Local Contrast Normalization (LCN) and ReLU to further elevate the model's capacity.

While the model performed well on both test and validation set, the assumption that the highest hand is always the target for the feature extraction for the sign language could not be always true this slight changes could introduce bias to the target gestures, this might also introduce errors. On this paper we utilized Media Pipe to focus on the most prominent hand through its parameter set to max complexity to note on this possible challenge prior to training.

Hand gesture recognition with depth images: A review [2] a paper that provides a comprehensive survey analyzing 37 research papers of hand gesture recognition that utilize depth cameras. The paper addresses three key queries; *What methods are being used to achieve hand localization and gesture recognition with depth cameras.* There are 10 common employed usage for hand tracking and gesture recognition namely, 2 segmentation, 3 for tracking gestures and 5 for classification. segmentation is done through depth thresholding, while for hand tracking methods include Kalman filters and mean shift, NITE body and hand-tracking within the OpenNI lastly models used for classification are Hiden Markov Models, KNN, ANN, SVM and Finite State Machines.

What applications and environments are researchers testing their methods in? Do they test them in situations where their depth-based methods have supposed advantages over video-based methods? Are the limitations of depth-based systems tested. Limited range of application are present for testing methods, minuscule work are done to leverage the advantages of depth cameras over intensity cameras for recognition of hand gestures.

How has the release of the Kinect and associated libraries affected research in gesture recognition? the advent of the use of Kinect has led to an increase in the number of research on depth-based gestures and also shift the goal of research in this area from classification to application, it was mentioned in the paper that among the 22 reviewed papers that utilized Kinect only 7 are for hand segmentation, 7 are for classification lastly 8 are for application with a notable increase in the

classification from the other 18 non-Kinect research where 13 papers are focused on tracking and 5 for tracking and segmentation.

While this research focuses on the classification without the use of Kinect, the paper tackles the first research question where utilizing MediaPipe with CNN on the palm images and Neural Networks to accomplish classification of hand gestures with tracking of hand coordinates.

III. Methodology

3.1 Data

The researcher used the data set from HAnd Gesture Recognition Image Dataset (HAGRID) [3]. Which consist of 18 classes of gestures but was used only for 3 gestures (Numerical 1, 3 and 4) in this research. Numerous volunteers totaling 37,853 contributed on the data set ranging from ages 18 to 65. The images are photograph on various conditions such are; volunteers facing and backing to a window, variations of lighting artificial and natural, images show gestures at distance of 0.5 to 4 meters from the camera. Lastly, hand gestures are made from 1 of either left or right hand with various positioning of Hand Landmarks. This research utilize 10,000 images from the data set per class total of 30,000 images with application of image preprocessing techniques before being fed to the models.

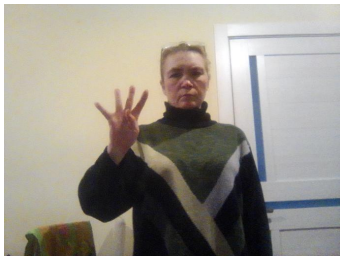
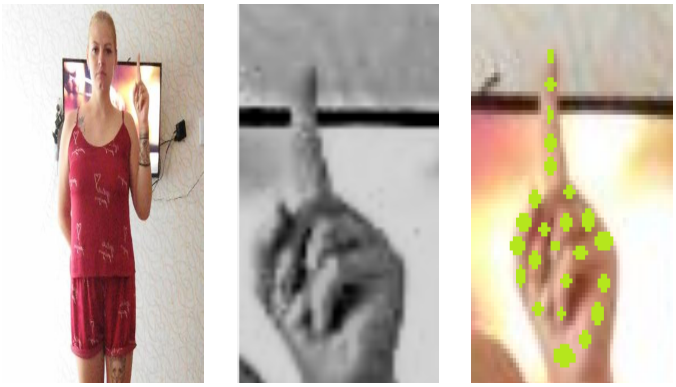


Fig. 1: Sample Right Hand Gestures of Four (4)

3.2 Data preprocessing

This research utilize two methods for identifying hand gestures, both utilize the MediaPipe library, MediaPipe is an open-source framework developed by google for building machine learning pipelines to process images, audio, and video. First is based on actual photograph resized to 28x28 and converted to grayscale then cropped hand that contains the gesture to be fed on the model.The 28x28x1 images consist of 784 features where each feature are colors of the images, flattened before being trained on the model.

Another, with MediaPipe library Hand Landmarks, it extracts the Hand Landmarks or referred in this research as coordinates, each blob means a feature of the photo (see Fig.2). These coordinates are generated from the distance of the root of the palm using a pre-trained model within the library to detect these coordinates. The MediaPipe hand coordinates resulted in 16 features with each feature consists of X and Y coordinates that are then flattened to be fed on the Neural Network model.



Original Photo Method 1 Method 2

Fig. 2: Various Methods

3.3 Neural Networks

Neural Networks inspired from human brain, is a machine learning algorithm that composed of interconnected nodes called neurons. Neurons received an input from the input layer and or hidden layers each neuron performs computation to produce an outcome.

Neural Network are compose of ”Nodes” where it computes the weighted sum of inputs and apply activation function that produces output for the next layer or predict the output of features. ”Layers” where compose of input layer where it receives the actual image or numerical representation of the image, hidden layer where features are extracted from the image and output layer where the model issues an answer based from the images’ class. **Weights and Biases** where weights are strength of each neuron while biases help adjust the output of the neuron. Lastly ”Activation Function” allows the model to recognize patterns and relationships in the data.

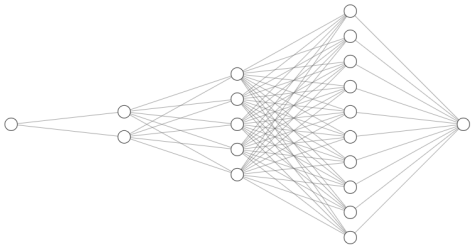


Fig. 3: Neural networks where the input is 1 feature with 3 hidden layers of [2, 5, 10] and 1 for the prediction

3.4 Convolutional Neural Networks

Convolutional Neural Networks is a class of machine learning algorithm like Neural Networks they are interconnected layers of neurons that processes the input data to come up with a prediction. Since Neural Networks struggle with high-dimensional data such as videos and images, CNN are designed to work on grid-like structures in images or videos.

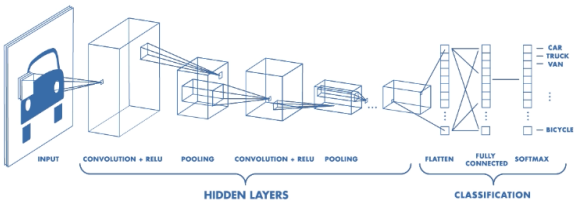


Fig. 4: Sample CNN model architecture

CNN has key components, the following are; **Convolutional Layers** that performs element-wise multiplication with overlapping region then sums the result to produce a single value in the feature map. **Pooling Layers** reduces the spatial dimensions of the input which decreases computation of the network while storing important feature. **Activation Functions** introduces non-linearity to the network, for it to learn complex patterns and relationship in the data. **Fully Connected Layer** last neuron of the network that connects all others layers to arrive with high-level reasoning.

IV. Experiments and Results

4.1 Experiments

The dataset for both models are set to 70% training set, 15% validation set and 15% test set, the researcher decided on this set-up due to the complex nature of images. Training set was batched due to the amount of the training set, batching allows to save time and reduce memory load. The models also utilize early stoppage and reduce learning rate to avoid overfitting of the models and each using Rectified Linear Unit (ReLU) activation.

Convolutional Neural Network used 9 layers with *first layer* input shape of 28x28x1 *Second layer* Conv2d of 64 filters 3x3, *third layer* max pooling of 2x2, *fourth layer* Conv2d of 128 filter 3x3, *fifth layer* max pooling of 2x2, *sixth layer* Conv2d of 256 filter 3x3, *seventh layer* max pooling of 2x2, *eight layer* condensed features to 32, lastly *ninth layer* of 3 features which are weights of each class between class 0 for “one”, class 1 for “three”, and class 2 for “four”.

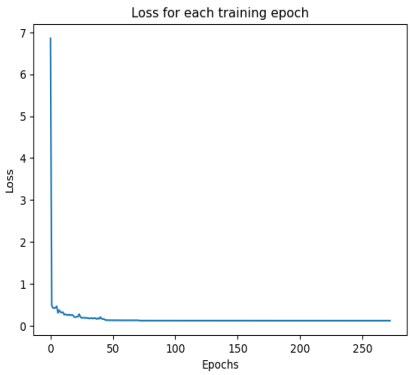
Conv2d (convolution layer) extracts the features from the 28x28x1 image, the pooling layers down samples the feature maps obtained from the Conv2d and Dense layer which flattens the features from the hidden layers through activation functions.

The second model utilizes Neural Network that has input size of 32 from the flatten X and Y coordinates of the hand landmarks with 3 hidden layers of 50, 100 and 255 layers using Adam optimizer trained on 200 epochs with reduce learning rate and early stoppage.

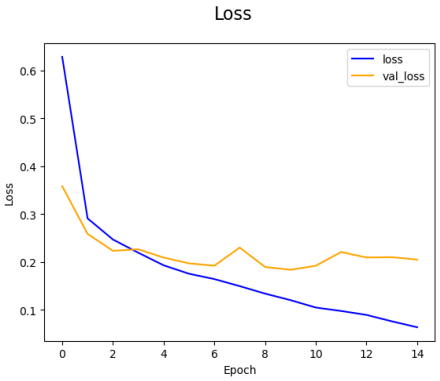
Notable training loss for both models are below 0.12 for Neural Networks and 0.20 for CNN model. (See Fig.5)

4.2 Results

Both models achieved excellent results, 92.47% on the test set for the model that utilized Convolutional Neural Networks



Loss on Neural Network Model



Loss on CNN model

Fig. 5: Training Loss

and 93.57% on test set for Neural Networks. What is intriguing in doing both models came from the processing time. The processing time was only tested on the actual training of the data set, for CNN model the training was between 350 to 478 seconds or between 6 to 8 minutes of training with only CPU, best time and accuracy at 406 seconds. The Neural Network model process took as low as 8.5 to 204 seconds of training with only CPU, best time at 95 seconds.

V. Conclusion

Both models strongly performed on the data set with respect to accuracy, the Neural Network model outperformed the CNN model thought not by huge margins (1%), and unexpectedly even with more complex computation within its model the Neural network required less time for processing the training with the data set. With the results in mind, the usage of MediaPipe hand landmark greatly increase the efficiency of models predicting images rather than feeding it with the actual images.

References

[1] Pigou, Lionel ,Dieleman, Sander, Kindermans, Pieter-Jan, Schrauwen, and Benjamin. (2015). Sign Language Recognition Using Convolutional Neural Networks. 8925. 572-578. 10.1007/978-3-319-16178-5_40.

[2] J. Suarez and R. R. Murphy, Hand gesture recognition with depth images: A review. 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, Paris, France, 2012, pp. 411-417, doi: 10.1109/ROMAN.2012.6343787.

[3] Kapitanov, A., Kvanchiani, K., Nagaev, A., Kraynov, R., and Makhliarchuk, A. (2024). HaGRID - HAnd Gesture Recognition Image Dataset. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 4572-4581)