Thai Sign Language Recognition: An Application of Deep Neural Network

Mayurdhvajsinh Jadeja : 92000133001
Pushti Depani : 92000133018

Research Paper Review

Abstract

Translating sign language is a challenging task due to intricacy of sign language structures including hand gestures, hand orientation, hand movements, and face emotion. Existing techniques mainly required complicated handmade elements to recognize a sign language. However, to develop a model based on those attributes is a challenging task. To deal with this difficulty, they suggested a model of Thai sign language recognition using Convolutional Neural Network (CNN) which can automatically learn both temporal and spatial information from data. This advantage is used to segment the hand sign in the colour video without the environment interference such as skin colour background. The histograms of oriented gradients are used to extract the image features of hand sign. These features are then pass to the artificial neural network for training and recognition. The result showed that the proposed method is robust to detect the hand gestures in the complex background.

Introduction

Thai Sign Language (TSL) is a visual natural language for connection or sharing information among people with hearing impairment. It conveys word or phrase meaning through hand gestures and alphabet. In Thailand, TSL is an efficient communication approach for impaired individuals in learning. Analysis of human movement and human-computer interaction HCI and the user's interface can boost the performance of Thai Sign Language Recognition. Moreover, computer can translate into text to increase interactivity between computer and human. In 2020, Thailand was found that 391,785 men from the total 2,076,313 or 18.87% are hearing impairments. It challenges to discover automatic TSL to support communicating with the hearing impaired and enabling the hearing impaired to interact with computers.

Sign language can be divided into two main categories: static and dynamic. Static signs are steady hand and face gestures, while dynamic signs are further divided into isolated signs and continuous signs. Isolated signs are hand gestures and facial expressions for a single word, i.e., ‘home’, ‘love’, and many more, while continuous signs are a sequence of isolated signs involving both hand gestures and facial expressions, i.e., ‘welcome to my new home’.

Convolution Neural Network

Convolutional Neural Networks (CNNs) is artificial neural networks used to extract feature and classify high information dimensional image. The main feature is extract feature and classify information directly.

There are mainly 4 layers in CNN:

Convolution Layer: Its purpose is to extract object feature by convolution value of each kernel which obtain from training. Each kernel is distinctly feature extraction. The first convolution layer obtained low level feature such as edges, lines, and angles etc. The next convolution layers will receive

Rectified Linear Units layer (ReLU layer). This layer has an activation function to decide which node is the next to send data.

Pooling layer. Subsampling data was occurred in this layer to decrease data variance and node in hidden layer.

Dropout layer. This layer is the hardest training layer because it has to dwindle down the repetitive remembering model by sampling probability value between 0.5-1.0 in each neuron.

The transformer layer takes the facial landmarks and the candidate face regions, then warp the face regions into a canonical pose by mapping the detected facial landmarks into a set of canonical positions. This explicitly eliminates the effect of rotation and scale variation according to the facial points.

The geometric transformation are uniquely determined by the facial landmarks and the canonical positions. In their cascade network, both the prediction of the facial landmarks and the canonical positions are learned in the end-to-end training process.

To make a final decision, we can concatenate the fine-grained feature from the second-stage RCNN network and the global feature from the first-stage RPN network. The concatenated features are then put into a fully connected layer to make the final face/non-face arbitration. This concludes the whole architecture of their proposed cascade network. The common practice is to train a prediction model to detect the facial landmarks, and then warp the face image to a canonical pose by mapping the facial landmarks to a set of manually specified canonical locations.

Recurrent Neural Network

RNN mainly focuses on the video, text data. Hand key points are saved in the form of csv file. RNN model consists of two recurrent neural networks. One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder–Decoder as an additional feature in the existing log-linear model. RNN can learn the temporal relationship of Speech – data & is capable of modelling time dependent phonemes.

MediaPipe

MediaPipe is the technology that provides the space tracking of the hands for this research using CNNs. The combination of space and time recognition allows the understanding of movement by a machine, which is what is needed to recognize sign language.

MediaPipe Framework

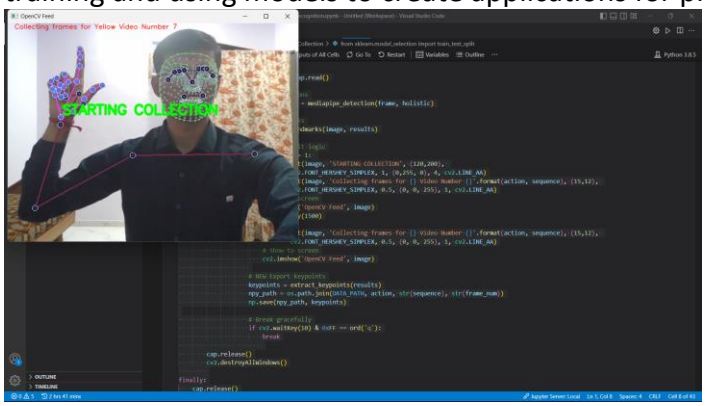
In this, the researchers used mediapipe for extracting hand key point that is already provided in the source code. MediaPipe is an open-source framework with a hybrid platform that creates pipelines for processing perceptual data, such as images, videos, and audio. It is an extensive approach employed with ML for hand tracking and gesture recognition in real-time.

Data preprocessing and feature extraction are carried over by the MediaPipe framework. Here, features from the face, hands, and body are extracted as keypoints and landmarks using built-in data augmentation techniques from sequence of input frames taken from a web camera.

In stage 2, the extracted keypoints from stage 1 are saved in a file to identify and remove the null entries from the data, after which data labelling follows. In stage 3, the cleaned and labelled gestures are trained and classified by our MOPGRU model for ISL recognition with the translated sign gestures in the form of text on the screen.

Approach

This can be done in two ways by extracting image and then concatenating it in the location. There are 4 steps: creating a video dataset, preparing data to train the model; Model training and using models to create applications for predict Thai sign language.



GRU and LSTM are used to overcome the problem of vanishing gradient. These methods are used to detect and identify sign language gestures from a video source and to generate the associated Thai word. foremost task is to divide the video file containing the sequence of ISL gestures for various words into separate sub-section videos containing different words. This is performed by identifying the start and end of each different gesture. After dividing up the videos, the next step is to divide the resulting subsection videos into frames.

The gesture recognition is further divided into two parts. The first one is static, and the second is dynamic. The pattern recognition problem belongs to the static gesture recognition, where the feature extraction is the part of preprocessing step. Feature extraction is an essential step in every conventional pattern recognition task. The static gestures require only a single image for processing input to the classifier, and it takes less computational cost.

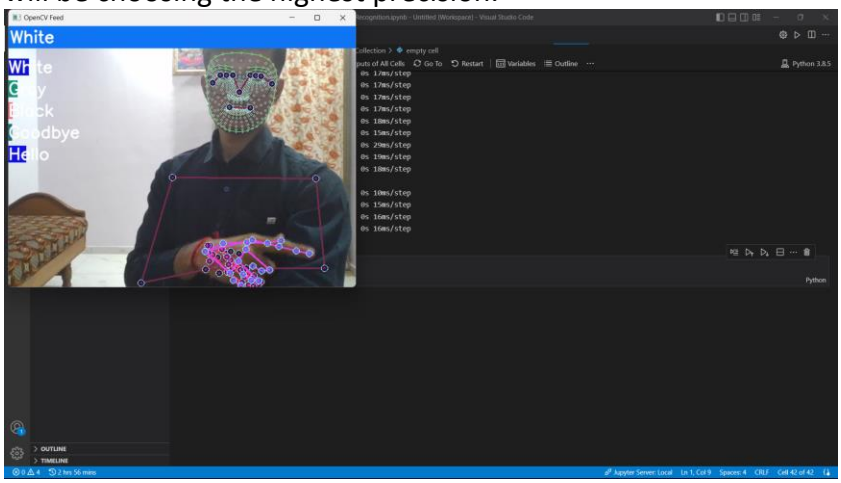
On the other hand, the dynamic gesture is the most challenging task in computer vision. It requires that a sequence of images and gestures are recognized based on features extracted from the proposed feature extraction algorithm. The deaf people mainly focus on learning the hand gestures for alphabets and digits to interact with others

Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning.

Input the data: Preparing the dataset for training and testing model by recording the video of the gestures.
Pre-processing: Loading the dataset using media pipe framework & extracting the hand key points and it will be written in the text file (csv). So to get accurate results we need to have the same size video to create equivalent dataset.

Model training : By using text file to train a model as it takes less time then he CNN approach. Also we need to repeat the process many times. Also we will be training the hyperparameter that will be focusing on the speed. Further it will be divided into nodes. Where we can conclude GRU had the least number of parameters while BiLSTM had the largest number of parameters, meaning BiLSTM was the largest model while the GRU was the smallest model in this experiment. Then we will be modelling the training data.

Post – processing: in this we will be building the desktop as well as mobile application. We will be choosing the highest precision.



Conclusion:
In this paper, researchers have proposed a simple model for Thai Sign Language recognition using media pipe. They have used the camera of laptop to capture the gestures using GPU. The movement changes detected by the model are: number of hands, vertical vs horizontal movement and vertical vs vertical movement. By increasing the number of layers in the LSTM and GRU, and applying LSTM followed by GRU, helps the model achieve higher accuracy in recognition of signs.