Modern University
for technology & information

Modern University
Faculty of
Computers
& Information

Hand Gesture Recognition

Mohamed Khaled Farahat
Ahmed Yousef Al Sayed
Ahmed Mohamed Ahmed Hussin
Gina Hamdy Kamal El-Din
Yousef Adel Mohamed

Dr. Hafez Abd El Wahab

T.A. Aya Khaled

Table of contents

Introduction

Human-Computer Interaction improvement using Computer Vision makes the computer more receptive to user needs. Physical Interaction with a computer today is not just limited to a keyboard and a mouse. Being able to interact with the system naturally using various sensory modes eases up the input process. Using computer vision in dynamic hand gestures like waving of hand while static hand gestures include joining the thumb and the forefinger to form the “Ok” symbol.

However, Computer Vision could be used in critical social applications to support physically challenged people. Computer Vision can convert sign language symbols to English characters. In this project, we will translate the sign language symbols to English alphabet characters using the Convolutional Neural Network (CNN). By training the CNN model the model detects hand symbols and predicts the correlated English letter through the sign language classification technique. Then we evaluate the model performance using the non-normalized and normalized confusion matrices to obtain the accuracy score of the CNN model.

Dataset

In this project, we used the American Sign Language (ASL) data set provided by MNIST which is publicly available at Kaggle. It contains 27455 training images and 7172 test images of size 28 x 28 pixels. These images belong to the 25 English alphabets starting from A to Y (Z is removed because it requires motion gesture). The dataset is available in the CSV format where training data has 27455 rows and 785 columns. The first column of the dataset represents the class label of the image and the remaining 784 columns represent the 28 x 28 pixels. same for the test data set. Then we covert our dataset matrix into a single vector CSV file using Flatten( ) function to ease up the navigation process and increase the performance of the model. After that, we input that CSV file into the preprocessing phase which extracts the main features of every hand gesture to be able to choose a correlated English character with the help of our labeled training data. That brings us to our last step where we evaluate our model using a test dataset.

Features

Features in CNN based models:

A CNN model can be thought of as a combination of two components: the feature extraction part and the classification part. The convolution + pooling layers perform feature extraction. For example, given an image, the convolution layer detects features such as finger numbers, the position of the fingers, the length of the finger a hand angle, etc. The fully connected layers then act as a classifier on top of these features and assign a probability of the input image is a character "A".

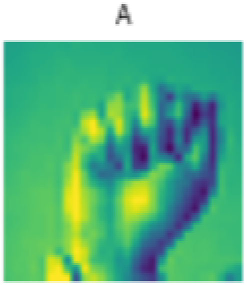


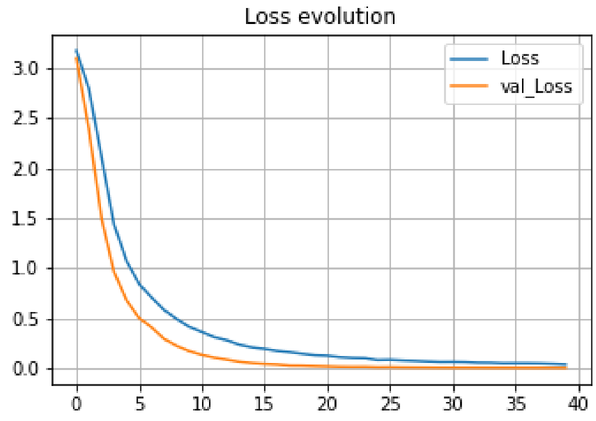
Fig Character "A"

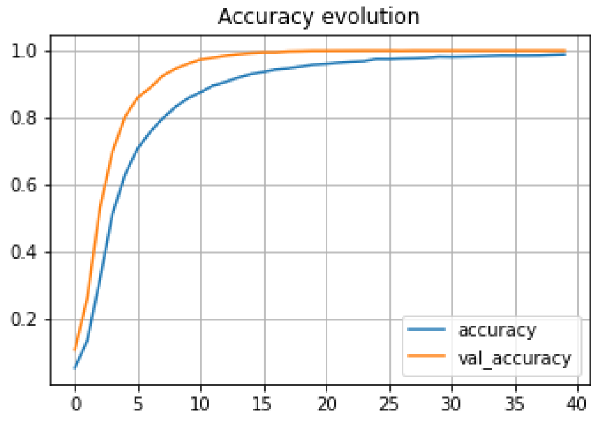
The convolution layers are the main powerhouse of a CNN model. Automatically detecting meaningful features given only an image and a label is not an easy task. The convolution layers learn such complex features by building on top of each other. The first layers detect edges, the next layers combine them to detect shapes, and to following layers merge this information to infer that this is a nose. To be clear, CNN doesn't know what a nose is. By seeing a lot of them in images, it learns to detect that as a feature. The fully connected layers learn how to use these features produced by convolutions in order to correctly classify the images.

Performance &Measurement Metrics

Measuring the performance of our model:

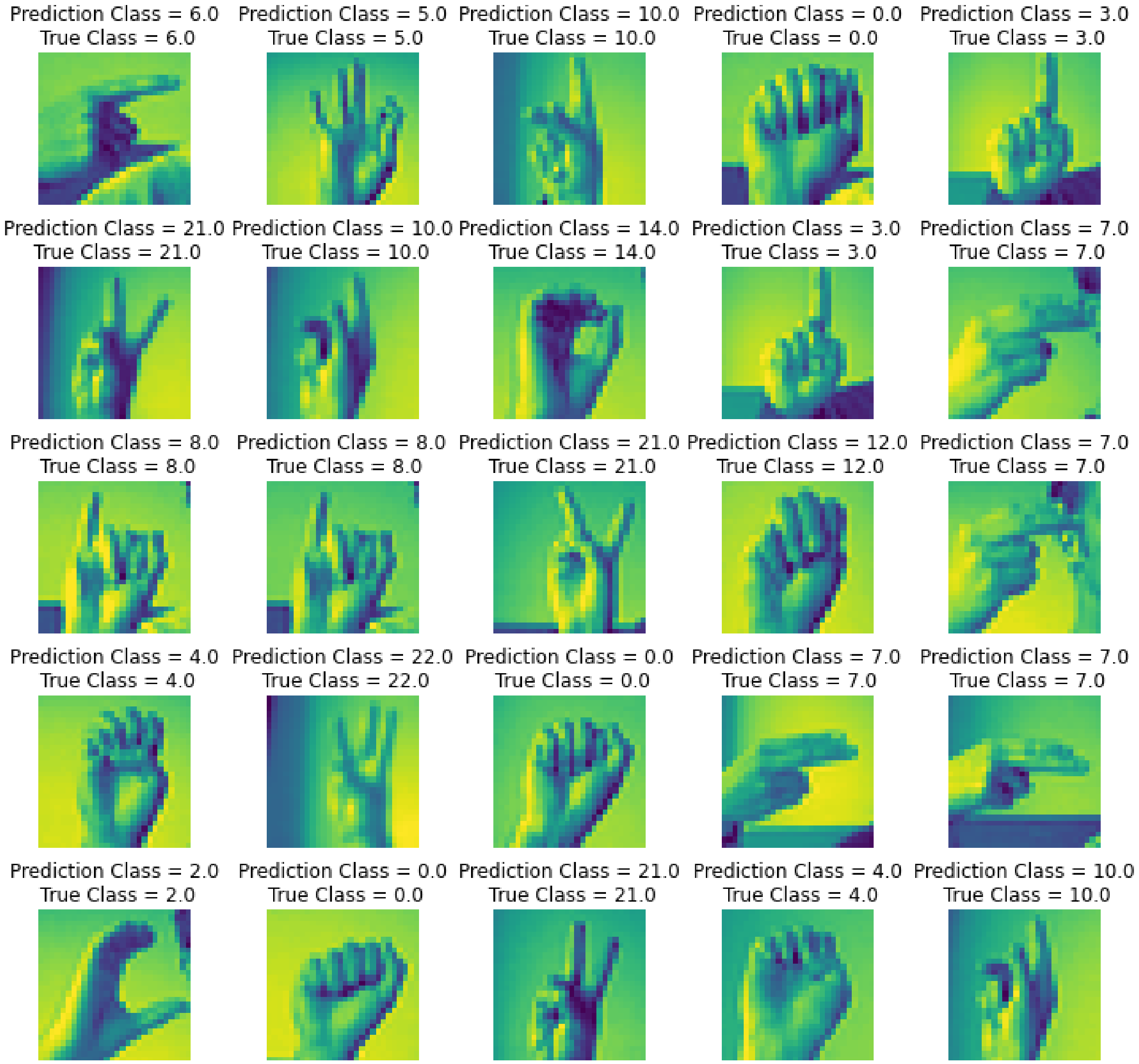
Here we can see that our model while comparing the loss evolution to the validation set evolution we can see that the curves are taking the intended shape this means that there is no overfitting or underfitting in our dataset and we can ensure that the prediction accuracy of our model will be very high and the loss evolution and the validation loss evolution are also very similar to the accuracy so we can expect a very low error rate since Loss is nothing but **a prediction error of Neural Net**. And the method to calculate the loss is called Loss Function. In simple words, the Loss is used to calculate the gradients. And gradients are used to update the weights of the Neural Net. This is how a Neural Net is trained.





accuracy checking:

Here we used a function to test the accuracy of the test set that we have and we got the following result:



Conclusion

A hand gesture-based recognition algorithm is proposed to control the PowerPoint application. In the proposed method, the foreground is extracted, through the Russian Mixture Model. The extracted object is applied to the Star Skeletonization process to detect the extreme points. The experimentation is tested on various datasets which justifies that the proposed solution outperforms the existing methods by being robust to scale variance and does not require any predefined templates for recognition.

Future work

The model could be re-engineered into a real-time detection model instead of static image detection to be able to translate words and full sentences not just one character at a time. Using a real-time model will enhance the usage of the project as a whole and make it more reliable for everyday use like voice input and dictation nowadays.