Hand Signals Classification Using CNN

Yin Nyein Kyaw
University of Technology
yinnyeinkyaw.ynk@gmail.com

Tin Myint Naing
Yatanarpon Cyber City
, utinmyintnaing08@gmail.com

Abstract – *Nowadays, driverless cars are very useful and effective for people. And so methods for controlling those driverless cars are needed and so hand signals recognition is very useful and effective way for that. Hand signals: Start, Stop, Emergency Stop and Slow Down play the important role for human and driverless cars interaction. This system uses Convolutional Neural Network (CNNs) for recognition of four kinds of hand signals- Start, Stop, Slow Down and Emergency Stop for controlling the driverless cars. A Convolutional Neural Network (CNN, or ConvNet) are a special kind of multi-layer neural networks, designed to recognize visual patterns directly from pixel images with minimal preprocessing. The input for the system is the hand-signal image of people from IP webcam and recognizes hand signals by using CNN-AlexNet Model. And the output is classification result which is one type of hand signal.*

Keywords – Hand Signals Recognition, Digital Image Processing, AlexNet Model

1. INTRODUCTION

Artificial intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. AI applications include expert systems, speech recognition and machine vision, visual perception, speech recognition, decision-making, and translation between languages [1]. Nowadays, Artificial Intelligence (AI) is a popular field and machine learning and deep learning are its subfield. Machine Learning enables IT systems to recognize pattern based on the existing algorithms and data sets and to develop adequate solution concepts. Machine Learning works in a similar way to human learning. A convolutional neural network (CNN) is a subfield of Artificial Intelligence (AI) too and it is widely used in pattern and image recognitions. That is specifically designed to process pixel data [3]. This system also uses CNN to recognize four types of hand-signals: Start, Stop, Emergency Stop and Slow Down for driverless cars. Recognition of hand signals is very important for users to control the driverless cars in easy ways.

2. MOTIVATION

This system focuses on the problem of hand signals recognition in real time that signal used by the users of driverless cars. The problem of hand signals recognition is based on the Digital Image Processing using Convolutional Neural Network (CNN-AlexNet Model). This system recognizes four types of hand signals: Start, Stop, Emergency Stop and Slow Down. This system aims to give users easy ways to control the driverless cars. This system can also save time and money for controlling the driverless cars in effective ways. But Convolutional Neural Network needs lots of training data to recognize the hand signals for getting the right and exact results. And so needed to collect the training data for the network. So the training data are the photos of people with four types of hand signals. There are about 400 training data for each hand signals: Start, Stop, Emergency Stop and Slow Down. Though, implementation of Neural Network is very simple but the accuracy can be based on the available numbers of training data for the network [5]. The main goal of this research paper is to demonstrate that how a good performance can be achieved without using any special hardware equipment and so that such a system can be implemented and easily used in real life. Hand Signals Classification (HSC) is the classification of human arms for four signal types: Start, Stop, Emergency Stop and Slow Down for controlling the driverless cars in easy and effective ways. For this process, people who will control the driverless car need to stand on the platform and make one of four hand-signal types that he or she want driverless car to do. And photo of that is taken by IP webcam to recognize what type of hand signal that is in the system.

3. METHODOLOGY

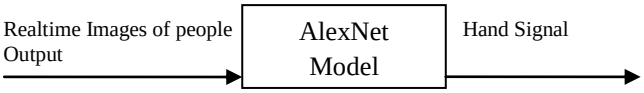

Figure 1. Overview of Hand Signal Classification

According to figure-1, this system will take the real time images of people from IP webcam to detect the types of hand signal with AlexNet Model. RGB images of size 227*227 are used as training images for AlexNet network. And the output of this system is one type of hand signal: Start, Stop, Emergency Stop or Slow Down.

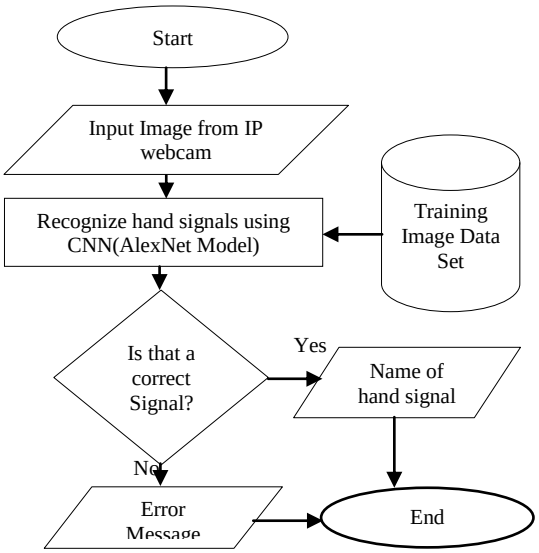

Figure 2. System Design of Hand Signal Recognition

- Firstly, photos of people (users of driverless cars) will be needed for the training network to recognize the types of hand signals.
- And the training data are RGB images of people who are standing on the platforms with different hand signals.
- So IP webcam camera is needed for taking the real time images of people.
- The IP address of that IP webcam is needed to use in MatLab coding to take photo.
- And Convolutional Neural Network (CNN-AlexNet Model) is used for hand signals recognition.
- That realtime photos is recognized in the AlexNet Model network step by step.
- And nearly all realtime photos of human hand signals: Start, Stop, Emergency Stop and Slow Down are recognized well with CNN-AlexNet Model.
- And so this system is very useful for the driverless cars users to control them in easy ways.

3.1. CNN Architecture

A Convolutional Neural Network (CNN, or ConvNet) are a special kind of multi-layer neural networks, designed to recognize visual patterns directly from pixel images with minimal preprocessing. Input layer has nothing to learn, at its core, what it does is just provide the input image's shape. So no learnable parameters here. Thus number of parameters = 0 [4]. The key components of a CNN are the convolutional layers. They are formed by grouping neurons in a rectangular grid. The training process aims to learn the parameters of the convolution. Number of parameters in a CONV layer would be : ((m * n)+1)*k), added 1 because of the bias term for each filter. The same expression can be written as follows: ((shape of width of the filter*shape of height of the filter+1)*number of filters) [5]. Pooling layers are usually placed after a single or a set of serial or parallel convolutional layers and take small rectangular blocks from the convolutional layer and subsample them to produce a single output from each block. This has got no learnable parameters because all it does is calculate a specific number, no backprop learning involved! Thus number of parameters = 0 [2]. Finally, dense layers (also known as "fully-connected" layers) perform classification using the features that have been extracted by the convolutional layers and then have been subsampled by the pooling layers. Every node of a dense layer is connected to all nodes of its previous layer. In mathematics, the convolution between two functions, say $f; g$: Rd -> R is defined as

$$[f \circledast g](x) = \int \mathrm{R}d\, f(z)g(x - z)dz.$$

That is, we measure the overlap between f and g when both functions are shifted by x and 'flipped'. Whenever we have discrete objects, the integral turns into a sum. For instance, for vectors defined on $\ell 2$, i.e., the set of square summable infinite dimensional vectors with index running over Z we obtain the following definition.

$$[f \circledast g](i) = \Sigma a\, f(a)g(i - a)$$

This diagram shows how convolutional neural networks combine layers that automatically learn features from many images to classify new images [4].

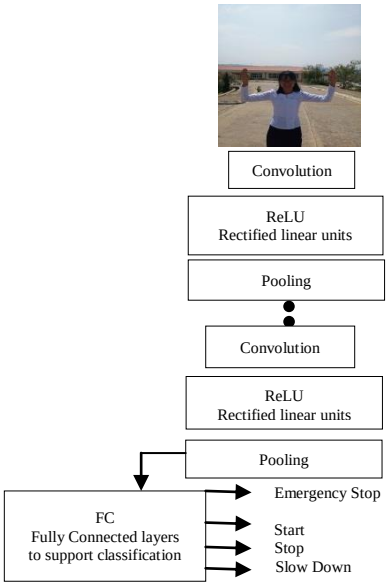

Figure 3. Layers Combination and Image Classification of CNN

In figure-3, the input image is detected according to the layers combination and image classification of CNN.

- First, the convolution operation extracts different features of the input image. The first convolution layer extracts low-level features like edges, lines and corners. Higher –level layers extract higher-level features.
- CNNs use many specific functions such as rectified linear unit (ReLU) to efficiently implement non-linear triggering. A ReLU implements the function y=max(x,0) and so this layer has same input and output sizes.
- The pooling or sub-sampling layer reduces the resolution of the features.
- Fully connected layers are used as final layers of CNN. And this layer mathematically sums a weighting of the previous layer of features.

3.2. AlexNet Model

AlexNet has been trained on over a million images and can classify images into 1000 object categories. The network has learned rich feature representations for a wide range of images. The network takes an image as input and outputs a label for the object in the image together with the probabilities for each of the object categories. It consisted 11x11, 5x5,3x3, convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum. It attached

ReLU activations after every convolutional and fully-connected layer [2].

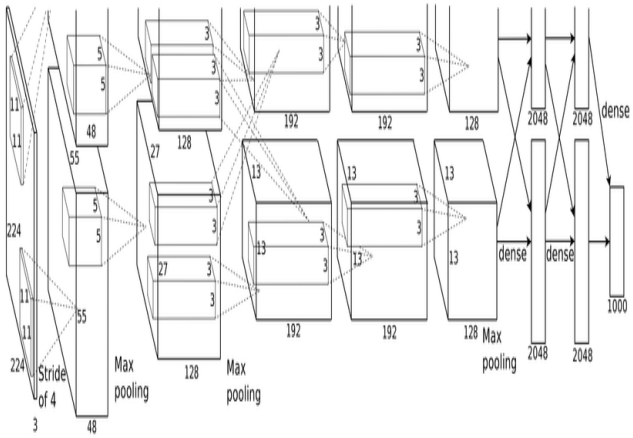

Figure 4. AlexNet Network

AlexNet architecture are**:** CONV1, MAX POOL1, NORM1, CONV2, MAX POOL2, NORM2, CONV3, CONV4, CONV5, Max POOL3, FC6, FC7, FC8.

Input: 227x227x3 images

First layer (CONV1): 96 11x11 filters applied at stride 4
The output volume size→ (227-11)/4+1 = 55
Output volume [55x55x96]
Total number of parameters for first layer: (11*11*3)*96 = 35K

After CONV1: 55x55x96
Second layer (POOL1): 3x3 filters applied at stride 2
The output volume size→ (55-3)/2+1 = 27
Output volume: 27x27x96
Total number of parameters for first layer: 0

Full (simplified) AlexNet architecture:
[227x227x3] INPUT
[55x55x96] CONV1: 96 11x11 filters at stride 4, pad 0
[27x27x96] MAX POOL1: 3x3 filters at stride 2
[27x27x96] NORM1: Normalization layer
[27x27x256] CONV2: 256 5x5 filters at stride 1, pad 2
[13x13x256] MAX POOL2: 3x3 filters at stride 2
[13x13x256] NORM2: Normalization layer
[13x13x384] CONV3: 384 3x3 filters at stride 1, pad 1
[13x13x384] CONV4: 384 3x3 filters at stride 1, pad 1
[13x13x256] CONV5: 256 3x3 filters at stride 1, pad 1
[6x6x256] MAX POOL3: 3x3 filters at stride 2
[4096] FC6: 4096 neurons
[4096] FC7: 4096 neurons
[1000] FC8: 1000 neurons (class scores) [2] [3].

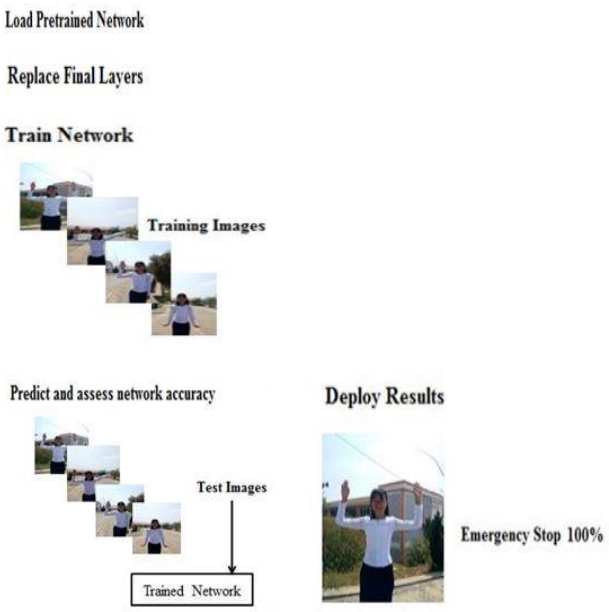

Figure 5. Reusing Pre-trained Network

In figure-5, this system loads the pre-trained AlexNet network and trains AlexNet with 1600 RGB training images of four types of hand signals with the size of 227*227: Start, Stop, Emergency Stop and Slow Down. And then testing images are used to test and assess the accuracy of trained network. And the testing result which is one type of hand signal is obtained.

3.3. Training Processing

To satisfy and reduce the computational effort needed for the processing, pre-processing of the image taken from the camera is highly important. Apart from that, numerous factors such as lights, environment, background of the image, hand and body position, parameters and focus of the camera impact the result dramatically. Training data will be needed to train AlexNet network to detect four hand-signals Start, Stop, Emergency Stop and Slow Down. For this purpose, RGB images of males and females with different types of clothes' colors with different backgrounds. Especially, most of the images are taken on the platforms. And there are about 400 training images for each type of hand-signal. And the following are some of the training images.

Figure 6. Training Images

And then AlexNet pre-trained network will be needed to load to train four types of hand-signals. And the following is the image of preprocessing or training.

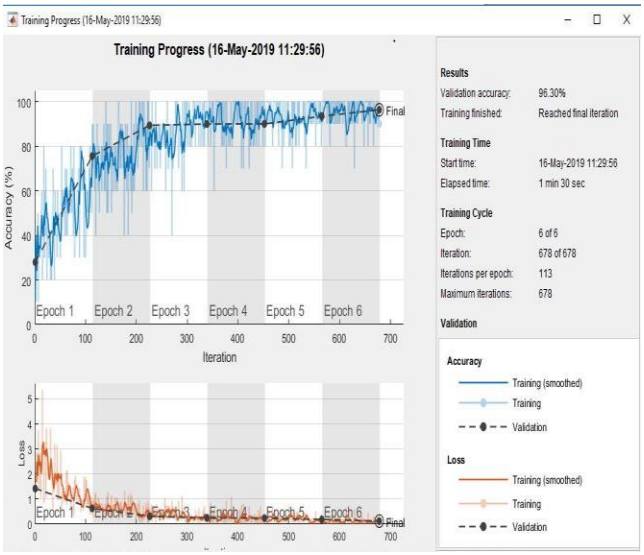

Figure 7. Training Progress

3.4. Testing Processing

In training network process, the validation accuracy is between 96% and 99%. And then, training images and validation images are divided into 70% and 30% respectively. And the following is the image for classification of validation images with training images.

Figure 8. Classification of Validation Images

3.4.1 Batch Images Testing

After training AlexNet network and classification of validation images and training images with percentages of 70 and 30, testing the network with the testing images to detect how trained network can work and recognize the hand-signals well. For this process, testing images are needed by taking new photos of the people which are not included in the training images. The clearer the background for hand-signal in the photo is, the higher testing accuracy can get more. Testing accuracy can be low if there are so many buildings and trees in the background. And the following are some of the images for batch images testing.

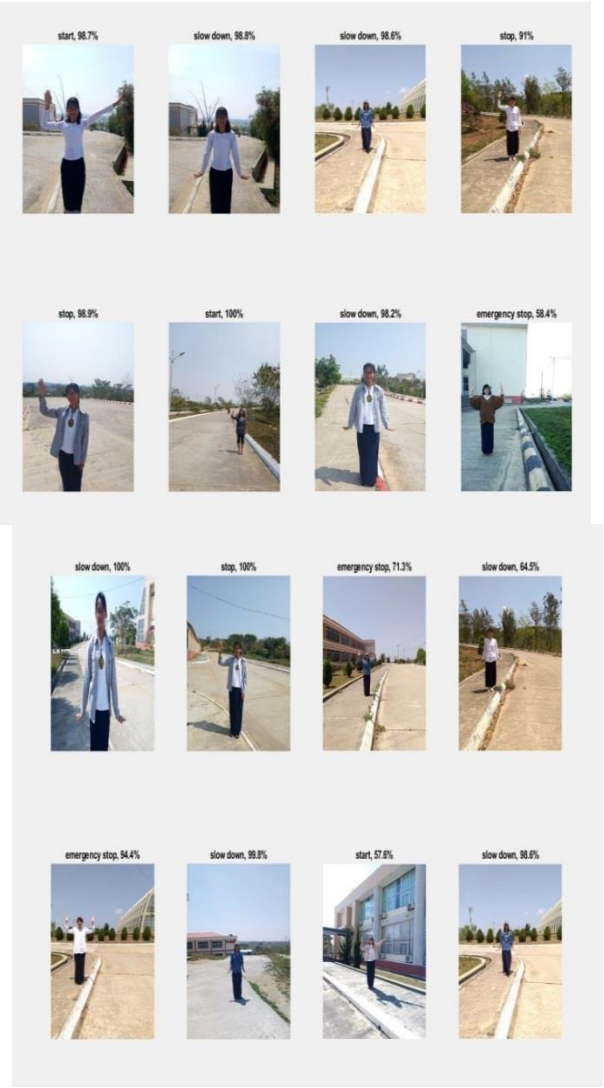

Figure 9. Batch Images Testing

3.4.2 Testing Results

There are 400 training images for each type of hand-signal and so 1600 images for total.

Table 1. Number of Predicted Hand-Signals

	Emergency Stop	Stop	Start	Slow Down
Emergency Stop(Total=400)	380	10	6	4
Stop(Total=400)	15	370	10	5
Start(Total=400)	5	10	380	5
Slow Down(Total=400)	7	5	3	385

Table 2. Accuracy for Each Hand-Signal Type

Accuracy	Emergency Stop	Stop	Start	Slow Down
	0.95	0.925	0.95	0.9625

4. CONCLUSIONS

This study is based on the problem of recognition four types of hand-signals: Start, Stop, Emergency Stop and Slow Down for driverless cars. And for this process, Convolutional Neural Network (CNN-AlexNet Model) is used to recognize the hand-signals. This AlexNet Model can recognize the type of hand-signals well nearly 100% for all testing images. And this system utilizes 400 training images for each type of hand-signal. And all these hand-signals are not difficult for the users of driverless cars. For testing results of this system, 'Slow Down' has the best testing result and 'Stop' has least accuracy. So the weakness of this system is less accuracy for complex background. So the future work will focus on background extraction and human detection for better detection of hand signals and higher accuracy.

5. ACKNOWLEDGEMENTS

The author offers her special thanks to Dr. Aung Win, Rector of University of Technology (Yatanarpon Cyber City), for his invaluable permission. The author would wish to record her thank to Dr. Hnin Aye Thant, Professor and Head, Department of Information and Communication Technology. And the author would like to express special and deepest thank to her supervisor, Dr. Tin Myint Naing, University of Technology (Yatanarpon Cyber City), for his guidance, suggestions and necessary advice. The author would like to express thank to all teachers from Department of Information and Communication Technology, University of Technology (Yatanarpon Cyber City) for their effective guidance and suggestions. Finally, the author's special thanks are sent to her parents for their supports, kindness and unconditional love.

6. REFERENCES

[1] May 2, 2017 – Fei-Fei Li & Johnson & Serena Yeung. Lecture 9- Today: CNN Architectures. Case Studies, AlexNet. VGG.

[2] "Hand Gesture Recognition using a Convolutional Neural Network", Institute of Informatics and Telecommunication National Center for Scientific Research .

[3] "MatConvNet. Convolutional Neural Networks for Anrea Vedaldi. Karel Lenc i. arXiv: 1412.4564v3 [cs.CV] 4 May 2016.

[4] "Multi-column Deep Neural Networks for Image Classification" Dan Ciresan, Ueli Meier and Jurgen Schmidhuber.

[5] "ImageNet Classification with Deep Convolutional Neural Networks", Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton.

[6] "Automatically identifying, counting and describing wild animals in camera-trap images with deep learning", Mohammed Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Ali Swanson, Meredith Palmer Craig Packer, and Jeff Clune.