Real-time Hand-Gesture Recognition based on Deep Neural Network

Naoto Ageishi*, *Fukuchi* Tomohide, and *Abderazek* Ben Abdallah

Adaptive Systems Laboratory, University of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu, Japan

Abstract. Hand gestures are a kind of nonverbal communication in which visible bodily actions are used to communicate important messages. Recently, hand gesture recognition has received significant attention from the research community for various applications, including advanced driver assistance systems, prosthetic, and robotic control. Therefore, accurate and fast classification of hand gesture is required. In this research, we created a deep neural network as the first step to develop a real-time camera-only hand gesture recognition system without electroencephalogram (EEG) signals. We present the system software architecture in a fair amount of details. The proposed system was able to recognize hand signs with an accuracy of 97.31%.

1 Introduction

Hand gestures are a form of nonverbal communication used by individuals in conjunction with speech to communicate. Nowadays, with the increasing use of technology, hand-gesture recognition is considered to be an essential aspect of Human-Machine Interaction (HMI), allowing the machine to capture and interpret the user's intent and respond accordingly. Hand-gesture recognition is a crucial task in medical settings such as prosthetic control.

A Neural Network (NN) is a type of machine learning method that mimics human neurons' activity. Neural network technology is used for the automatic driving of cars, games, monitoring, and voice recognition [1–7]. Among them, in the gesture recognition system using neural networks, its application using various tools and algorithms, and its importance are attracting attention in robot control [8].

Gesture recognition is recognized using various tools depending on the purpose. Accurate gesture recognition can be achieved by introducing new acquisition devices such as Leap Motion and Kinect [9]. As one of the gesture recognition tools, there is one called the Electroencephalogram (EEG) signal. An EEG signal is a sensor attached to the head that reads a signal from the brain trying to move one's arm and move the artificial limb from that signal. Gesture recognition using only EEG signals cannot obtain high accuracy, but gesture recognition using multiple devices can get higher accuracy than one device [10].

This research will develop a camera-only hand gesture recognition system using Convolutional Neural Network (CNN) for research that initially requires two sensors, an EEG signal, and a camera, to perform gesture recognition. We will create a real-time system that recognizes ten types of American hand sign numbers 1-10.

The structure of this thesis is as follows. Section 2 shows the proposed system architecture, American hand sign datasets, and convolutional neural network. Section 3 shows results of the proposed American hand sign recognition system. Section 4 and 5 describe discussion and conclusion.

*Corresponding Author: Naoto Ageishi e-mail: s1250022@u-aizu.ac.jp

2 System Overview

Figure 1 shows the system overview of hand gesture recognition. In Figure 1, one frame of the image from the camera was cropped. The cropped image is preprocessed. Then, the cropped image was used for the learned Deep Neural Network (DNN). The image's pose is the output. It then shows what gesture the identified image represents.

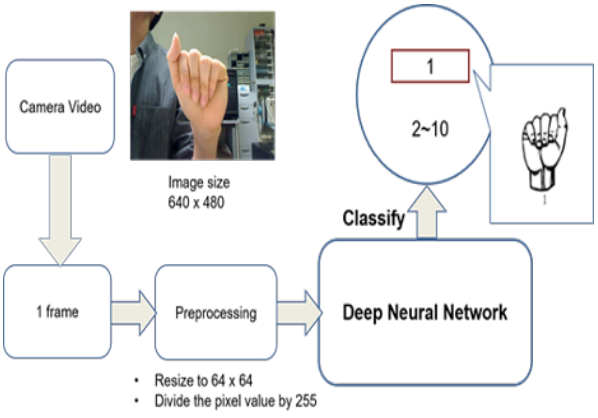

Figure 1. System Overview of Hand Gesture Recognition

2.1 American Hand Sign Dataset

In order to collect data, we first put our elbows up in front of the video camera and then recorded our hands moving

closer to and away from the camera. We also changed the background and shot a total of 23 types of videos.
We converted the video files we recorded into frame images and cut out 150 images for one gesture. Since 10 kinds of gesture poses, as shown in Figure 2, were taken in one video, the number of images was 1500 in one video. 5 of the 20 videos were shot by one volunteer and the remaining 15 were shot by Ageishi. These were divided into training and validation; 16 types of videos were used for training data, and 4 types were used for validation data. As shown in Table 1, 150 images were extracted from one video for each gesture. A total of 24,000 images of 16 types of videos were used as the training data, and a total of 6000 images of 4 types of videos were used as the validation data. As the test dataset, we used three test videos taken by two people. As the test dataset, we used three test videos taken by two people. These individuals were not duplicated in the training and validation datasets.

Table 1. Number of dataset of training and validation

Gesture	1	2	3	4	5	6	7	8	9	10	Total
Training images	2400	2400	2400	2400	2400	2400	2400	2400	2400	2400	24000
Validation images	600	600	600	600	600	600	600	600	600	600	6000

2.2 Image Preprocessing

Preprocessed the image before training and inferring on the model. The images were resized from 640 x 480 to 64 x 64. They were divided by 255 to be normalized and the pixel values were set in the range 0.0-1.0. Created the correct label for the gesture. Both the normalized images and the correct labels were saved in npy format.

2.3 Convolution Neural Network

The CNN model was trained based on the model shown in Figure 3, and the results of the 24,000 training data are as shown in Figure 4. The model included convolutional neural network layers, max-pooling layers, dropout layers and fully connected layers. As the activation function, the Relu function was used for the convolutional neural network layer and fully connected layers. Softmax function was used for the output of the accuracy. Dropout was used to prevent overfitting, and the learning of the model proceeded smoothly by using these.
Since this DNN started overfitting, we performed early stopping at 21 epochs. DNN learning results showed that the accuracy of validation data was 97.31 %, and the loss value was 0.1312.

2.4 Training and Execution Environment

The training environment of our convolutional neural network and execution environment are as follows. We performed measurements with 480p and 30 fps settings using the camera QCAM-200SX. Deep learning framework Keras 2.4.3 was used.

Table 2. Model accuracy and fps when changing image size

Input image size	Accuracy (%)	fps
28 x 28	96.07	19
32 x 32	96.00	18
64 x 64	97.31	18
128 x 128	96.20	17

3 Evalaution Results

Table 2 shows the accuracy of the training model by resizing the images used for training. Highly accurate models were created for each image size. Input image size of 64 x 64 had the highest accuracy.
Figure 5 shows the results of the recognition accuracy of the test videos with the trained model. This confusion matrix shows which gesture the model was for the pose gesture. The bottom row shows the correct gesture label, and the left row shows the model inference results. Coloured cells show the number of images that the model identified correctly. Although there were some mistakes, American hand sign commands were recognized highly accurately. The execution speed of identification for the captured data set was about 18 fps, which was fast, but the real-time identification with the camera connected was slower than the camera setting of 30 fps.
Figure 6 is an example of real-time identification using a camera. When we take one of the ten poses in front of the camera, the image is cut into one frame, resized and normalized to be identified in the trained model. The result is displayed in the upper left corner of a window called a frame.

4 Discussion

In this research, we created a real-time American hand sign recognition system. As shown in Table 2, the size of the input image was changed and learning was performed. Among them, the size of 64 x 64 was 1.11% higher than the size of the other input images in terms of accuracy. We attempted the learning with an image size larger than 64x64, but the results were not very good even if the image size was quite large. As well, if the size of the image was too large, the size of the model also became enormous, and it took a long time to infer. Large images were not suitable for real-time recognition. The 64x64 input images have a slightly slower fps than other small input sizes, but in this study, we prioritized obtaining a higher accuracy probability and used it for inference.
Although high accuracy was obtained in the test, the gesture could not be recognized correctly depending on the amount of light at the measurement location. Since learning and measurement were performed with only CNN, recognition may not always be correct depending on the surrounding shooting environment. In order to solve these problems, it is necessary to create a system that identifies the position of the fingers.
The strength of the system using only CNN was that fps was not so low, which could enable a real-time recogni-

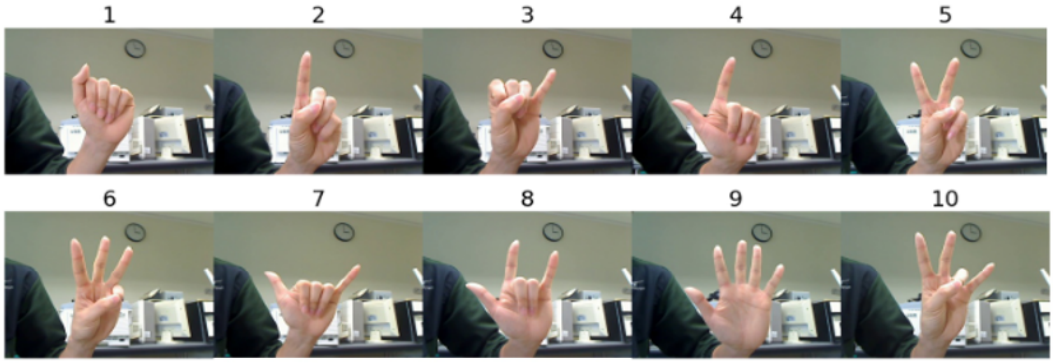

Figure 2. Frame images of American Hand Gesture 1-10 cut out from video

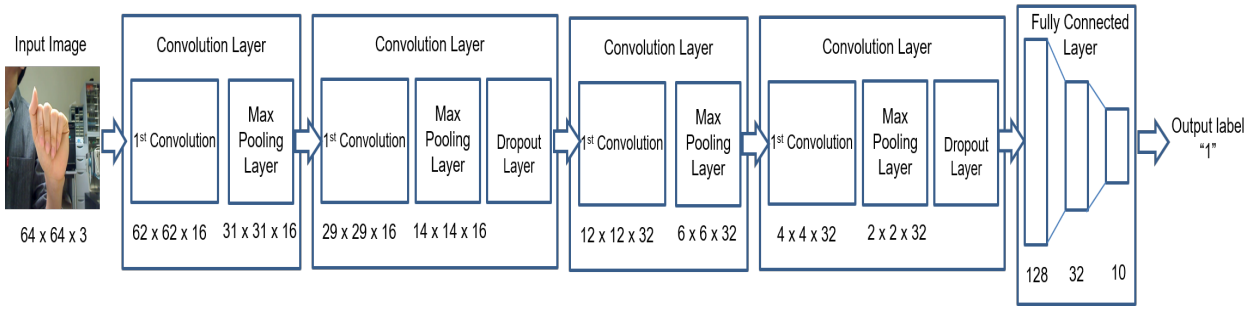

Figure 3. Structure of Deep Neural Network

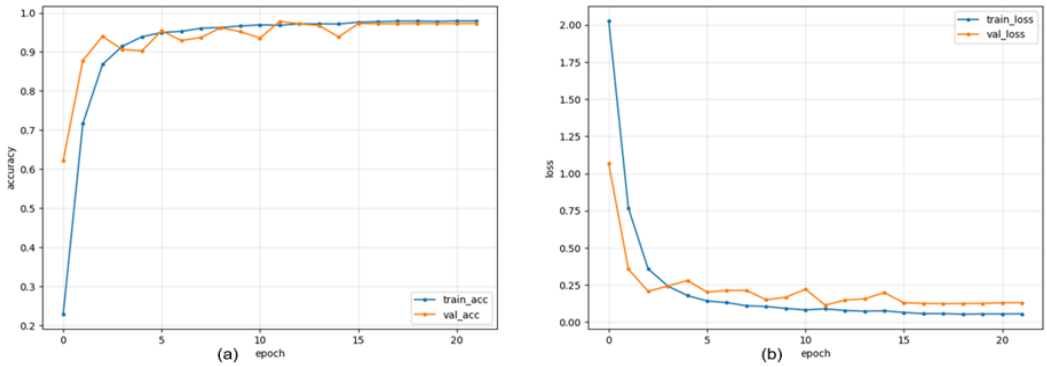

Figure 4. Learning results, (a): Accuracy of learning result, (b): Loss value of learning result

tion system. However, for hardware and GPU implementation, even faster speeds are required for sensor fusion; hardware can be executed at high execution speed and low power consumption.

5 Conclusion and Future Work

In this study, we proposed a real-time hand sign recognition system using DNN. We were able to recognize 10 types of gestures with high accuracy. However, the recognition accuracy varied depending on the shooting environment.

As a future activity toward sensor fusion, we will implement it in hardware so that it can be executed at higher speed and lower power consumption.

References

[1] T.H. Vu, R. Murakami, Y. Okuyama, A.B. Abdallah, *"Efficient Optimization and Hardware Acceleration of CNNs towards the Design of a Scalable Neuro inspired Architecture in Hardware"*, in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)* (2018), pp. 326–332

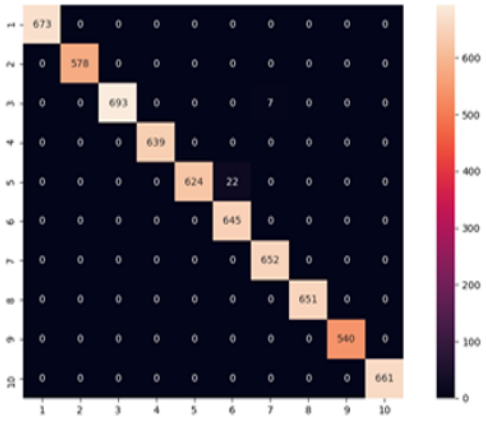

Figure 5. Confusion matrix of test movie recognition

Figure 6. Example of inference from the camera

[2] Y.O. T. H. Vu, A.B. Abdallah, *"Comprehensive analytic performance assessment and k-means based multicast routing algorithm and architecture for 3d-NoC of spiking neurons"*, in *ACM Journal on Emerging Technologies in Computing Systems, vol. 15, no. 4* (Dec. 2019. [Online]), p. 1–28, https://doi.org/10.1145/3340963

[3] O.M.I. T. H. Vu, A.B. Abdallah, *"Fault-tolerant spike routing algorithm and architecture for three dimensional NoC-based neuromorphic systems"*, in *IEEE Access, vol. 7, 2019. [Online]* (2019. [Online]), p. pp. 90 436–90 452, https://doi.org/10.1109/access.2019.2925085

[4] R. Murarami, Y. Okuyama, A.B. Abdallah, *"Animal Recognition and Identification with Deep Convolutional Neural Networks for farm Monitoring"*, in *Information Processing Society Tohoku Branch Conference, Koriyama, Japan* (Feb,10,2018)

[5] Y. Murakami, Y. Okuyama, A.B. Abdallah, *"SRAM Based Neural Network System for Traffic Light Recognition in Autonomous Vehicles"*, in *Information Processing Society Tohoku Branch Conference, Koriyama, Japan* (Feb. 10, 2018)

[6] T. Fukuchi, M.O. Ikechukwu, A.B. Abdallah, *"Design and Optimization of a Deep Neural Network Architecture for Traffic Light Detection"* (EDP Sciences, 2020), Vol. 77, p. 01002, https://doi.org/10.1051/shsconf/20207701002

[7] K.D. M. Ogbodo, T. Vu, A. Abdallah, *"Light-weight spiking neuron processing core for large-scale 3d-NoC based spiking neural network processing systems"*, in *in 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), IEEE* (Feb. 2020. [Online]), https://doi.org/10.1109/bigcomp48618.2020.00-86

[8] R.Z. Khan, N.A. Ibraheem, *"Hand Gesture Recognition: A Literature Review"*, in *International Journal of Artificial Intelligence Applications (IJAIA), Vol.3, No.4, July 2012* (2012), pp. 1–14

[9] G. Marin, F. Dominio, P. Zanuttigh, *"Hand gesture recognition with Leap Motion and Kinect devices"*, in *IEEE International Conference on Image Processing (ICIP), Paris, France, 2014* (2014), pp. 1–5

[10] E. Ceolini, C. Frenkel, S.B. Shrestha, G. Taverni, L. Khacef, M. Payvand, E. Donati, *"Hand-Gesture Recognition Based on EMG and Event-Based CameraSensor Fusion: A Benchmark inNeuromorphic Computing"* (2020), pp. 1–15