Hand Gesture Classification based on Inaudible Sound using Convolutional Neural Network

Jinhyuck Kim[1], Jeongung Kim[1] and Sunwoong Choi[2]

[1]Department of Secured Smart Electric Vehicle, Kookmin University of Seoul, Korea.
[2] School of Electrical Engineering, Kookmin University of Seoul, Korea.

Abstract. Recognizing and classifying the gesture of a user has become important for an increase in the use of wearable devices. This study propose a method for classifying hand gestures by creating inaudible sound using a smartphone and reflected sound signal. The proposed method converts the sound data, which has been reflected and recorded, into an image using short-time Fourier transform (STFT), and the obtained data are applied to a convolutional neural network (CNN) model to classify hand gestures. The results showed classification accuracy for 6 hand gestures with an average of 92.17%.Furthermore, it is confirmed that the proposed method has a higher classification accuracy than other machine learning classification algorithms.

1 Introduction

As IoT products grow and wearable devices become more popular, more and more research is being done on ways to recognize human behavior and gestures. There are studies that recognize gesture using various sensor or parts such as optical sensor[1] and radio frequency chip[2]. In another study, studies were conducted to recognize human behavior and gestures using sound waves without additional components[3-4].

In this paper, we propose classification of hand gestures using the reflection effect of sound waves. Using a smart phone, the sound of the non-audible range that can not be heard by the human ear is generated, and the reflected signal is recorded and collected. The STFT is applied to the recorded signal and the image is classified and classified through the CNN model to evaluate the accuracy. As a result of the experiment, we showed classification accuracy of 92.17% for 6 kinds of hand gesture.

2 Proposed Method

First, we make our own applications to collect sound data reflected by hand gestures. The application performs the function of recording and generating a single inaudible frequency of 20 kHz for 3 seconds. Using two smartphones, one plays a speaker and one plays a microphone, collecting sound reflected by the hand gesture.

Next, the collected data is preprocessed so that it can be applied to the classification model. The recorded sound data is imaged using STFT. The 19.8 kHz to 20 kHz section corresponding to the non-audible frequency range of the imaged data is cut out. It also cuts off the start 0.2 seconds to eliminate the system delay errors that occur internally during the recording process. As a result,

we will use 2.8 seconds of data corresponding to 19.8 kHz to 20 kHz for each hand gesture. Figure 1 shows the recorded raw data as a graph and the STFT applied data as a spectrogram.

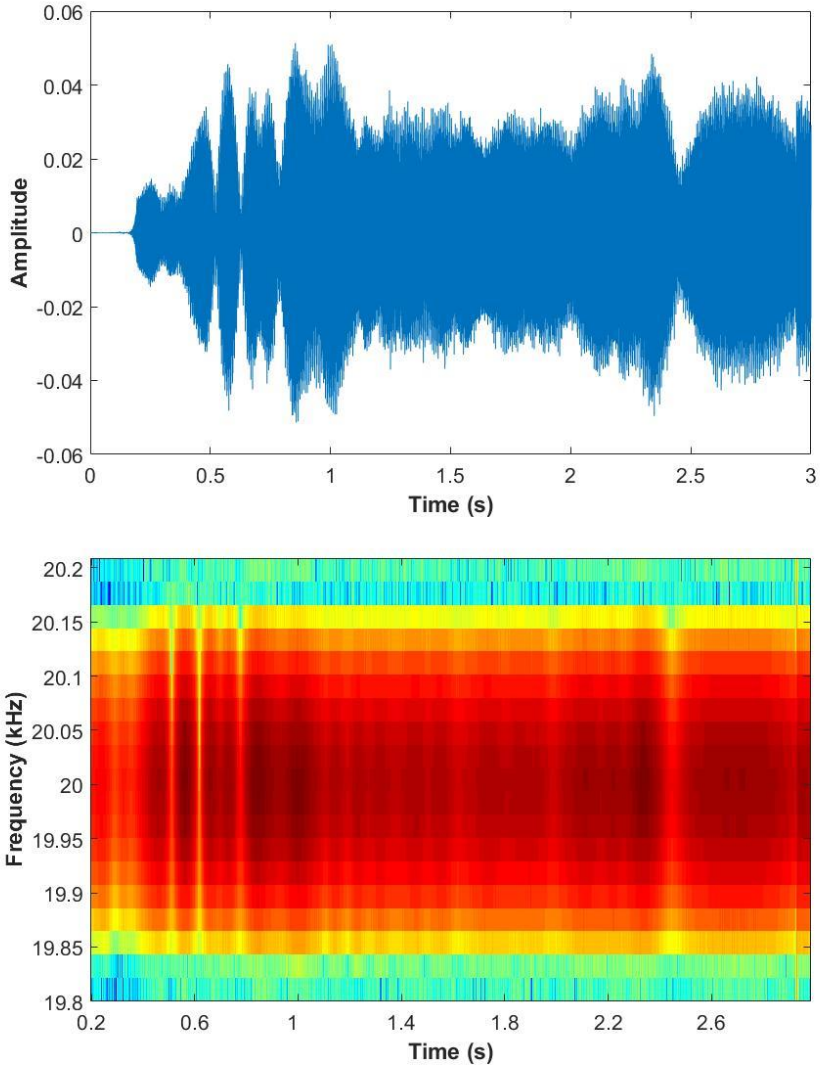

Fig. 1. Recorded raw data graph (above) and Spectrogram with STFT (below)

Finally, the preprocessed data is learned by the CNN model implemented in the server and the classification performance is evaluated. The proposed CNN model for classification is a 9 - layered model. As input, it receives 20x4920x1 size data, and it passes through each convolution layer to reduce size through Max pooling. After passing through the last convolution layer, average pooling is applied after Max pooling. Finally, the average pooling result is adjusted to the target number of labels using the Fully Connected layer and the prediction results are obtained using the Softmax function. Figure 2 shows the proposed CNN model.

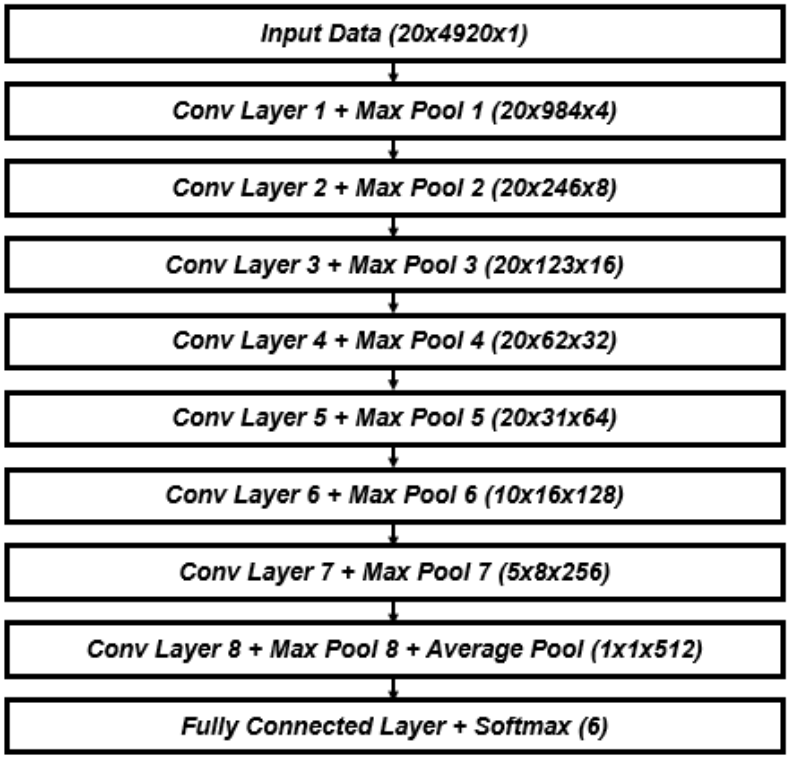

Fig. 2. The structure of the proposed CNN model

3 Evaluation

3.1 Experiment Condition

In this paper, we classify the six hand movements. The hand gestures to classify are as follows. (1) No action is taken during recording. (2) Move your hand from left to right while holding your palm while recording is in progress. At this time, the hand is placed on the screen of the smartphone to record. (3) Move the hand from right to left while holding the palm while recording is in progress. At this time, the hand is placed on the screen of the smartphone to record. (4) While the recording is in progress, draw a circle clockwise from the index finger. At this time, the finger is placed on the screen of the smartphone to record. (5) While the recording is in progress, use the palm of your hand to stop the microphone at the bottom of the smartphone. (6) Move your hand from right to left while holding your palm while recording is in progress. At this time, the hand is placed on the screen of the smartphone to record. We collected 100 data for each hand operation and preprocessed it. For the evaluation using CNN model, the whole data was divided into 8:2 and used as learning data and evaluation data.

3.2 Experiment Result

As a result of the experiment, we showed classification accuracy of 92.17% for 6 kinds of hand movements. For the performance comparison, we confirmed the classification accuracy using SVM (Support Vector Machine) and RF(Random Forest), which are machine learning classification

models with the same data. As a result, the SVM model showed 73.3% classification accuracy and the RF model showed 78% classification accuracy for the 6 kinds of hand gestures.

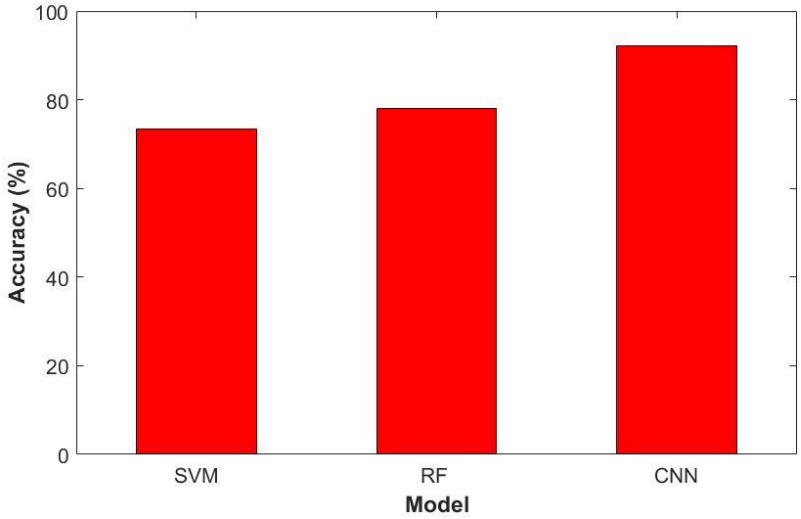

Fig 3. Comparison of hand gesture classification accuracy about each model

4 Conclusion

In this paper, human hand gestures are classified using inaudible frequencies. We recorded the sound reflected from the smartphone using the developed application, and applied the STFT to image the data. After that, data were learned in the CNN model and the six kinds of hand movements were classified. As a result, the proposed method showed classification accuracy of 92.17% and showed higher accuracy of classification of hand movements when compared with machine learning algorithms such as SVM and RF. In the future, we will carry out research to improve the performance of classification models by adding hand motions and using data collected in other environments.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) Grant funded by the Korean Government(MSIP) (No.2016R1A5A1012966).

References

1. Lien, J. Gillian, N. Karagozler, M. E. Amilhood, P. Schwesig, C. Olson, E. Raja, H. and Poupyrev, I.: 'Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar', *ACM Trans*, 35, 2016
2. Zhang, C. Tabor, J. Zhang, J. and Zhang, X.: 'Extending Mobile Interaction Through Near-Field Visible Light Sensing', *MobiCom*, 345-357, 2015
3. Gupta, S. Morris, D. Patel, S. N. and Tan, D.: 'SoundWave: Using the Doppler Effect to Sense

Gestures', *CHI*, 1911-1914, 2012
4. Gao, H. Xu, X. Yu, J. Chen, Y. Zhu, Y. Xue, G. and Li, M.: 'ER: Early Recognition of Inattentive Driving Leveraging Audio Devices on Smartphones', *INFOCOM*, 2017