Hand Gesture Classification using Non-Audible Sound

Jinhyuck Kim, Jinwon Cheon
Dept. of Security Enhanced Smart Electric Vehicle
Kookmin University
Seoul, Korea 02707
{j0910, jinwontoo}@kookmin.ac.kr

Sunwoong Choi
School of Electrical Engineering
Kookmin University
Seoul, Korea 02707
schoi@kookmin.ac.kr

***Abstract*—**Recognizing and distinguishing the behavior and gesture of a user has become important owing to an increase in the use of wearable devices, such as a smartwatch. This study aims to propose a method for classifying hand gestures by creating sound in the non-audible frequency range using a smartphone and reflected signal. The proposed method converts the sound data, which has been reflected and recorded, into an image within a short time using short-time Fourier transform, and the obtained data are applied to a convolutional neural network (CNN) model to classify hand gestures. The results showed classification accuracy for 8 hand gestures with an average of 87.75%. Additionally, it is confirmed that the suggested method has a higher classification accuracy than other machine learning classification algorithms.

Keywords—non-audible sound; hand gesture; gesture classification; convolutional neural network; short-time fourier transform

I. 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 87.75% for 6 kinds of hand gesture.

II. RELATED WORK

Various studies that integrated sound wave into IT technology have been conducted. The phase shift of sound wave was used to track finger positions. Strata [5] conducted a study on tracking finger positions using sound wave. In a study that used a low-latency acoustic phase (LLAP) [6], a phase shift of sound wave was converted into the length of an object's movement to track finger positions. A static vector and a dynamic vector were calculated to measure changes and find the position of the finger.

Lastly, there are studies that classify behaviors or gestures using sound wave. ER [7] classified 4 behaviors in a car using a sound wave. It generated sound waves using a smartphone mounted inside of a car, and extracted characteristics of each behavior based on the Doppler effect. Then, the behaviors were classified using principal component analysis and SVM. Resultantly, it showed a classification accuracy of 94.8% on 4 behaviors.

III. PROPOSED METHOD

Figure 1 shows the system architecture of the suggested method. First, we collect sound data reflected from hand gestures using the proposed application. The application comprises a function that generates a non-audible frequency for a certain period of time and a function that records sound. Two smartphones were used, one being a speaker and another being a microphone. A single band non-audible frequency of 20 kHz was generated through the speaker built in a smartphone, and the generated signals were recorded by the microphone built in another smartphone. While the smartphone was recording, we were able to perform each gesture and obtain each different signal reflected accordingly.

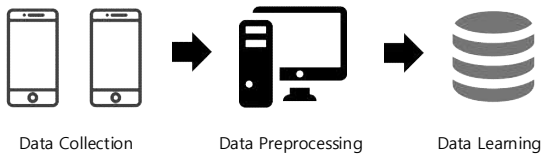


Fig. 1. Overall system architecture.

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, use 2.8 seconds of data corresponding to 19.8 kHz to 20 kHz for each hand gesture. Figure 2 shows the recorded raw data as a graph and the STFT applied data as a spectrogram.

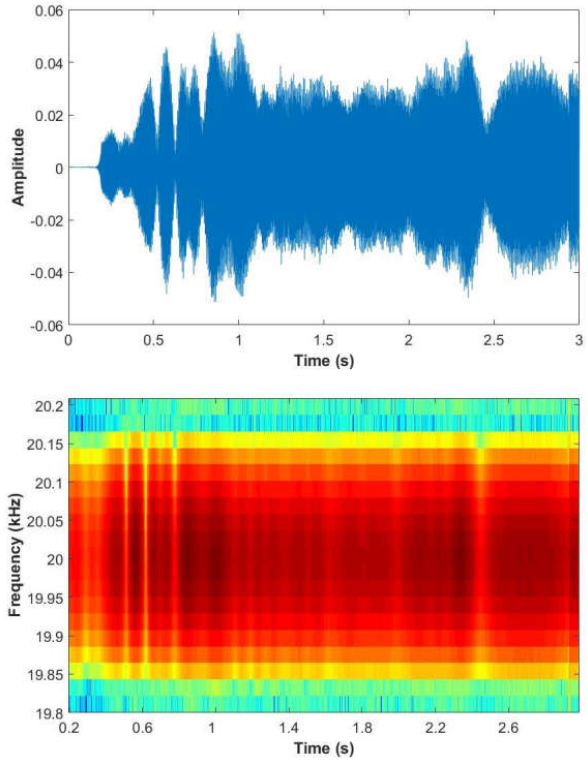


Fig. 2. Recorded raw data graph and spectrogram after STFT.

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 3 shows the proposed CNN model.

IV. Evaluation

A. Experiment Condition

Recorded data per direct behavior were collected to evaluate the performance of the proposed method. For data collection, we made an application that generates and records a single non-audible frequency of 20 kHz. Figure 4 (a) is the UI of the application.

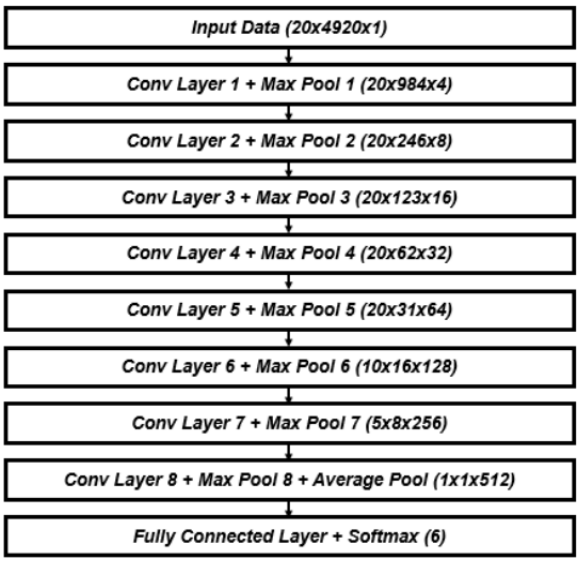
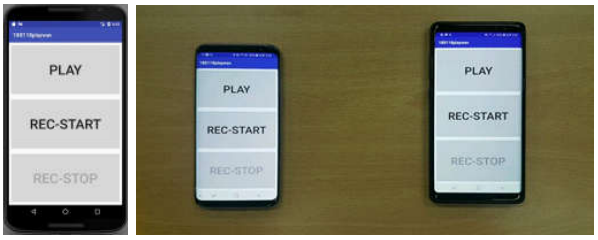


Fig. 3. The structure of the proposed CNN model.

We used 2 smartphones for this experiment. Samsung Galaxy S8 model was used as a speaker that generates non-audible frequencies, and Samsung Galaxy Note 8 model was used as a microphone for recording sounds. The application was installed on each smartphone to proceed with the experiment.



(a) (b)

Fig. 4. (a) UI of application for collecting data; (b) Experiment environment for collecting actual data.

The experiment was performed in a deserted laboratory. Two smartphones were placed on a table with space and the PLAY button is pushed on the smartphone that works as a speaker. Then, the REC-START button is pushed on the smartphone that works as a microphone before taking particular hand gestures. Data is collected by repeating this process. Figure 4 (b) shows an experimental environment for collecting actual data.

A GPU server was used to implement and test the CNN models. The GPU server used the GTX 1080 Ti model. With Tensorflow, we implemented 2 CNN models to learn and evaluate data before and after applying STFT.

B. Gesture Dataset

In this paper, we classify the six hand movements. Figure 5 illustrates hand gestures collected for the experiment. Each hand gesture was completed within 3 seconds. We collected data by repeating 8 hand gestures as illustrated. The hand did not touch the smartphone screen while making hand gestures and was positioned about 1 cm away from the screen. 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.

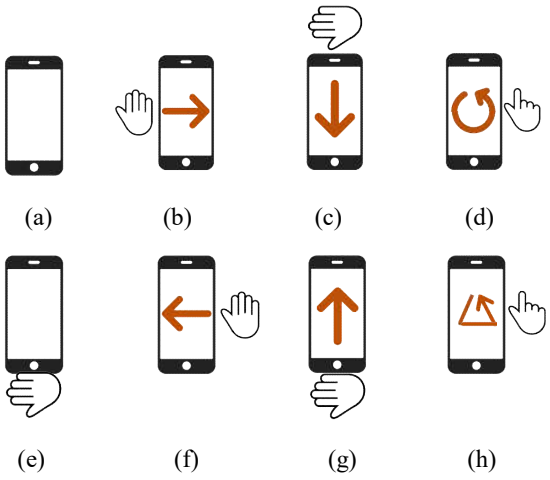


Fig. 5. Used Gestures (a) Do nothing; (b) Move from left to right; (c) Move from top to bottom; (d) Circle drawing; (e) Block the microphone; (f) Move from right to left; (g) Move from bottom to top; and (h) Triangle drawing.

C. Experiment Result

As a result of the experiment, we showed classification accuracy of 87.75% for 8 kinds of hand movements. For the performance comparison, we confirmed the classification accuracy using DT (Decision Tree), SVM (Support Vector Machine) and RF (Random Forest), which are machine learning classification models with the same data. As a result, the DT model showed 49.63% classification accuracy, the SVM model showed 71.25% classification accuracy and the RF model showed 79.63% classification accuracy for the 8 kinds of hand gestures. Figure 6 compares the accuracy of 3 machine learning algorithms and the suggested CNN models.

V. 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 87.75% and showed higher accuracy of classification of hand movements when compared with

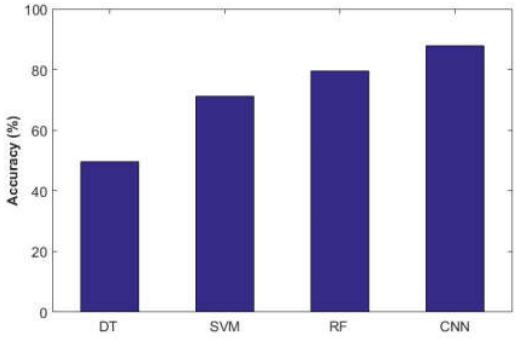


Fig. 6. Comparison of hand gesture classification accuracy about each model.

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 *(Heading 5)*

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.

[5] Yun, S.; Chen, Y. C.; Zheng, H.; Qiu, L.; Mao, W. Strata: Fine-Grained Acoustic-based Device-Free Tracking. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys), Niagara Falls, New York, USA, 19-23 June 2017; pp. 15-28.

[6] Wang, W.; Liu, A. X.; Sun, K.; Device-Free Gesture Tracking Using Acoustic Signals. In Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking (MobiCom), New York, USA, 03-07 October 2016; pp. 82-94.

[7] Gao, H.; Xu, X.; Yu, J.; Chen, Y.; Zhu, Y.; Xue, G.; Li, M. ER: Early Recognition of Inattentive Driving Leveraging Audio Devices on Smartphones. In Proceedings of the IEEE Conference on Computer Communications (INFOCOM), Atlanta, GA, USA, 01-04 May 2017.