ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 967-971 (2020)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Hand Gesture Classification using Inaudible Sound with Ensemble Method

Jinwon Cheon[1], Sunwoong Choi[*,2]

[1]*Department of Security Enhanced Smart Air Mobility, Kookmin University, Seoul, 02707, South Korea*

[2]*School of Electrical Engineering, Kookmin University, Seoul, 02707, South Korea*

ARTICLE INFO

Article history:
Received: 07 September, 2020
Accepted: 02 November, 2020
Online: 08 December, 2020

Keywords:
Hand gesture classification
Short-Time Fourier Transform (STFT)
Ensemble

ABSTRACT

Recognizing the human behavior and gesture has become important due to the increasing use of wearable devices. This study classifies hand gestures by creating sound in the inaudible frequency range from a smartphone and analyzing the reflected signals. We convert the sound using Short-Time Fourier Transform to magnitude and phase. We trained two types of data on Convolutional Neural Network model. And then we propose a method applying soft voting, an ensemble technique, to improve classification accuracy taking the average of two models' result. In this paper, the classification accuracy of the Mag model is 96.0% and the classification accuracy of the Phase model is 90.0% for 8 hand gestures. While the ensemble model showed 96.88%, which is better than Mag and Phase models.

1 Introduction

With the advancement of IT technology, the use of wearable devices such as smartwatches and IoT-based devices is becoming more popular. However, as devices become smaller for portability, there are some limitations, including the difficulty of controlling these devices using buttons or touches. To date, lots of studies are being conducted to overcome this limitation. Most studies collect and recognize data from sensors like cameras and controllers [1]–[4]. However, it is a fatal flaw that requires a sensor or product to recognize movement. Therefore, in this study, we propose a hand gesture classification using a smartphone without a separate sensor.

When the smartphone's speaker makes an inaudible sound, the microphone records this sound. A subject performs a hand gesture while recording is in progress, the signal hit by the subject's hand and reflected changes in frequency domain because of the Doppler effect. There are several methods to analyze a signal into time and frequency domain, among them, we analyzed the signal using the Short-Time Fourier Transform (STFT) [5]. This is because STFT converts the data into a 2D grid-like topology with time and frequency domains. It can be easy to input in 2D Convolutional Neural Network (CNN) [6]. In this way, the sound signal is transformed into magnitude and phase. We propose a method to increase the classification accuracy of the independently trained magnitude and phase CNN models by applying soft voting, an ensemble technique. As a result, the proposed method showed a classification accuracy of 96.88% for 8 hand gestures.

The rest of this paper is organized as follows. It introduces the hand recognition and classification research using ultrasound and ensemble technique in Chapter 2, describes the method proposed in Chapter 3, evaluates its performance in Chapter 4, and finally concludes this paper in Chapter 5.

2 Related Work

In this chapter, we briefly introduce how hand gesture recognition and classification are being used, and studies using ultrasonic signals and studies to improve the performance of a model using multiple CNN models are described.

First, the current status of research on hand gesture recognition and classification is introduced [7]. Data for hand gesture recognition and classification are collected from various sensors, for example, mount-based sensors, multi-touch screen sensors, vision-based sensors, and ultrasonic-based sensors. The collected data are trained on machine learning models used in various fields, such as human-robot interaction, virtual manipulation, sign language, and gaming.

Second, studies on recognizing or classifying hand motions using ultrasonic signals will be described. There is a study called

[*]Corresponding Author: Sunwoong Choi, Kookmin University, School of Electrical Engineering, 82-2-910-4416 & schoi@kookmin.ac.kr

FingerIO that uses smartphone and smartwatch to find fingers using the microphone and speaker [8]. In this study, they use sound in a inaudible frequency range and track the fingers' position by Orthogonal Frequency Division Multiplexing (OFDM). Other study proposes a deviceless gesture tracking scheme. This is called LLAP (Low-Latency Acoustic Phase) [9]. In this study, they use microphones and speakers that built-in on mobile devices to track hands and fingers. This study measures the phase changes that is formed by movement and converts the changes into the distance of the movement. There is a paper that classifies hand gestures using two smartphones [10]. One used as a microphone, while the other used as a speaker. In this paper, they used STFT. These window size is 500, and overlapped size is 475. The frequency resolution is 2,048. As a result, their proposed CNN model classified 8 hand gestures with a classification accuracy of 87.75%.

Finally, studies using multiple CNN models will be described. These two papers [11, 12] explained ensemble method. In paper [12], the performance of the model for classifying hand gestures was improved using the ensemble technique. They ensembled three types of classifiers: a polynomial classifier, a multi-layer perceptron, and a support vector regression. So, they obtained median recognition rates per moment in time of more than 86%. Another paper [13] is a paper that improves performance by fusion of GoogLeNet, VGGNet, and ResNet. The performance was improved by extracting feature layers (100-dimensional and 40-dimensional fully connected and softmax) to each model and attaching them. This paper implemented a model that recognizes the motions of a person in an image, and performance improved in order when only one of the three models mentioned above was used, when two were used, and when all three were used.

3 Proposed Method

Our system can be divided into three steps (data acquisition, data pre-processing and model training) like Figure 1. First, we collect sound data using our own application. Then, the data are pre-processed by STFT on MATLAB. Lastly, We train the pre-processed data on the proposed CNN model.

Figure 1: Overall system architecture.

3.1 Data Acquisition

We collected data with a smartphone application developed in Android Studio. This application produces a single frequency of 20 kHz on the smartphone's speaker. At the same time, the application starts recording on the smartphone's microphone. The speaker and microphone are activated for 3 seconds, and a subject makes a hand gesture within that time at a distance of 1–5 cm from the smartphone. The application stores the sound reflected by gestures. When collecting data, only one subject in the office performing hand gestures to reduce the noise.

3.2 Data Pre-processing

We have used STFT to convert sound data because STFT informs the change of frequency over time. A window size of 5,000, an overlap size of 4,750, and a frequency resolution of 2,048 are detailed information of STFT we used. The microphone recorded for 3 seconds, but the speaker only produced up to 2 seconds. So, we cut off the last second of the data. We also trimmed the data around 20 kHz to suppress the background noise and reduce the number of data. As shown in the Figure 2, the raw data graph represents amplitude value only, however the STFT spectrogram represents the signal intensity using color. STFT data contain magnitude and phase parts. Figure 3 shows magnitude of STFT data, and Figure 4 shows phase of STFT data. So, we pre-processed the sound data to magnitude and phase data.

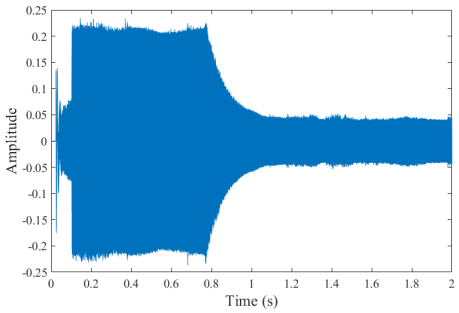

Figure 2: Recorded raw data amplitude graph.

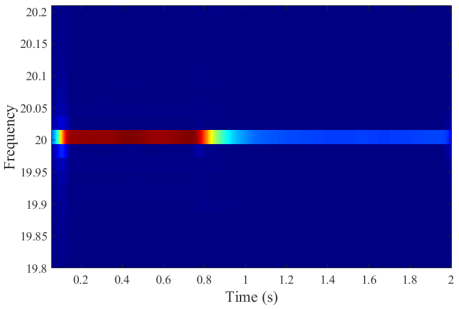

Figure 3: STFT spectrogram (magnitude).

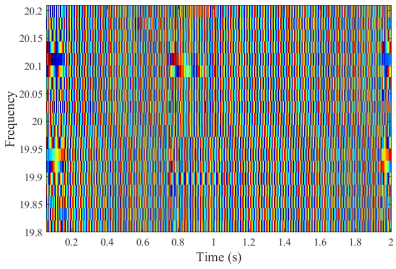

Figure 4: STFT spectrogram (phase).

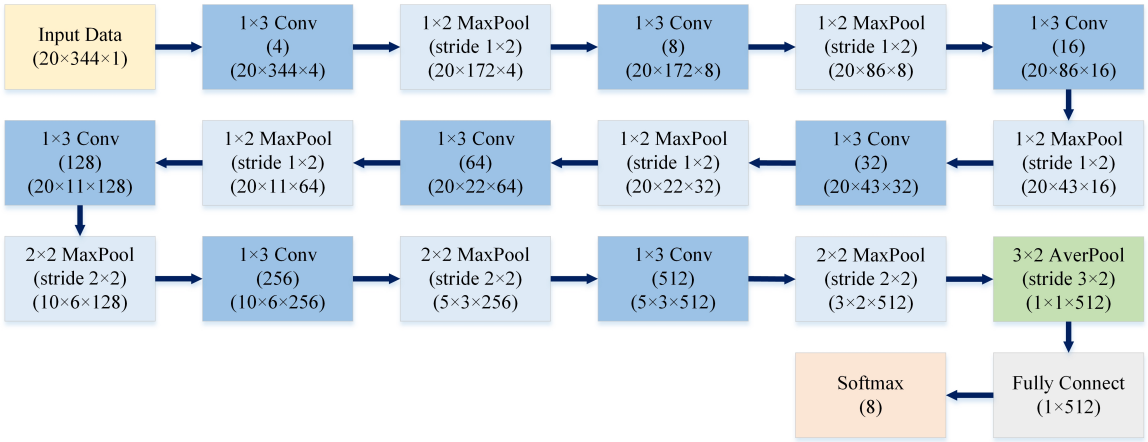

Figure 5: The structure of Mag and Phase models.

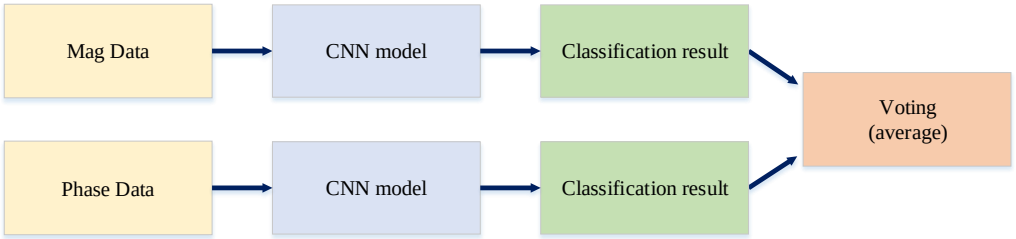

Figure 6: The structure of ensemble model.

3.3 Model Training

We propose two CNN models which called Magnitude (Mag) model and Phase model. Mag and Phase models have the same architecture. The models trained with one part of pre-processed data (magnitude or phase). The data were trimmed to the specific frequency range of 19.8–20.2 kHz through STFT. The models are largely composed of 5 parts: convolution layers, max pool layers, an average pool layer, a fully connected layer and a softmax function. There are 8 pairs of a convolutional layer and a max pool layer in the models to extract data feature. The filter size of the convolutional layer is 1×3, and the filter size of the max pool layer is 1×2 and 2×2. The output of pairs of the convolutional layer and the max pool layer passes through an average pool layer of 3×2 and then through a fully connected layer. Figure 5 shows the structure of our CNN models.

After that, the models were ensembled by taking the average of the softmax function outputs (soft voting). When Mag and Phase models train, the ensemble model not affected training but is only used when testing. Two results of softmax function are calculated by inputting magnitude data and phase data for one hand gesture into each model. The size of the softmax result of each model is the same as the number of classes, and a new softmax result is created by averaging two softmax results corresponding to each hand gesture. The operation is as (1). In this equation, S means the result of softmax funcion.

$$S^{ensemble} = \frac{S^{mag} + S^{phase}}{2} \quad (1)$$

Figure 6 shows how to progress the ensemble model. The magnitude part of the STFT is input to the Mag model and the phase part is input to the Phase model to calculate the average of the softmax matrix outputs from each model. The class that has the largest value among the values of the average matrix is the predicted class of the ensemble model.

4 Evaluation

4.1 Experiment Setup

In the experiment, we have collected the data to TG&Co.'s LUNA smartphone in an office where only one subject performed hand gestures. The application was developed in the Android Studio environment. The UI of developed application is as Figure 7. The smartphone was placed on a flat desk and recording started when the REC-START button was pressed. The application activated the speaker to produce a single frequency at 20 kHz and the microphone to start recording. After 3 seconds, recording stopped automatically.

The saved sound data were transferred from the smartphone to the computer. The data were converted to STFT on MATLAB R2019b. The models were implemented using the Tensorflow 1.6.0, and Python 3.4.8 for running the Tensorflow. Training and test were performed on a GPU server which operates CentOS 7.5 and uses Intel Xeon Silver 4114 CPU. The learning progressed about 4,000 epochs. Besides, a 5-fold cross validation technique was used to increase the reliability of the models.

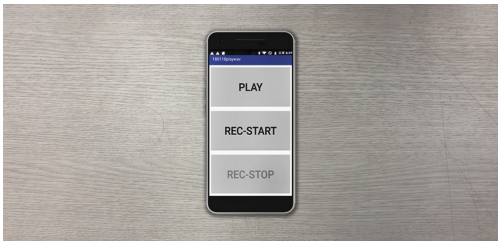

Figure 7: UI of developed application.

4.2 Gesture Dataset

In this study, 8 hand gestures were classified. Figure 8 shows 8 hand gestures that were used as a dataset. The subject gestured at a distance of 1–5 cm from the phone without touching the screen. This action was performed within 3 seconds and repeated 100 times per gesture. Through this process, we have collected a total of 800 data. The data were divided into training data and test data at a ratio of 8:2.

- Do nothing : Do not move the hand while recording.
- Block : Block the microphone with palm while recording.
- Move : With the palm open, record perpendicular to the screen. Place the hand at the beginning of the arrow and move it in the direction of the arrow while recording (to right, to bottom, to left, to top).
- Draw : Stretch only index finger and draw on the screen once in the direction of the arrow (circle and triangle).

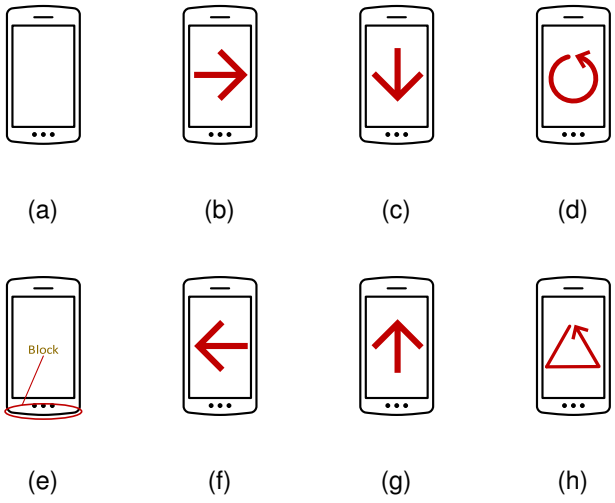

Figure 8: Used hand gestures.

4.3 Experiment Result

When performing STFT, the STFT result value also changes depending on the window size, so that we evaluated this. We tested the classification accuracy by increasing the window size from 2,000 to 8,000 in steps of 1000 using only magnitude data. In all window sizes, the frequency resolution and overlap size were unified to 2,048 and 95% during the evaluation. The classification accuracy by window size is shown in Figure 9. Therefore, the evaluations of the Phase and ensemble model were proceeded when window size is 5,000.

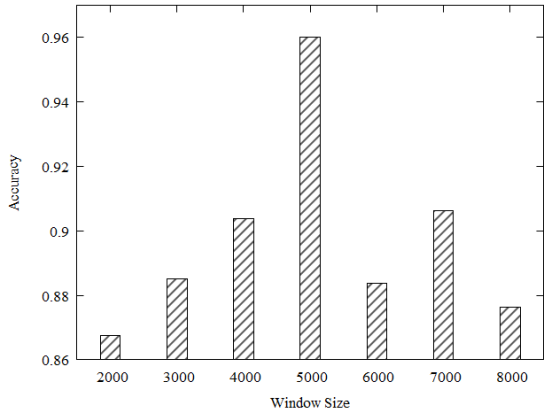

Figure 9: Classification accuracy by window size.

In conclusion, we presented the ensemble model that classified 8 hand gestures as 96.88% classification accuracy. We estimated Mag, Phase, and the ensemble model when window size is 5,000. Mag model showed 96.0% classification accuracy, and Phase model showed 90.0%, slightly lower than Mag model. However, when these two models were ensembled with our method, the performance improved, showing a classification accuracy of 96.88%. Not only classification accuracy but also various evaluation indicators were tested. Figure 10 shows precision, recall, F1-score, and accuracy of each model. It can be seen that the ensemble model better than the two models in all evaluation indicators.

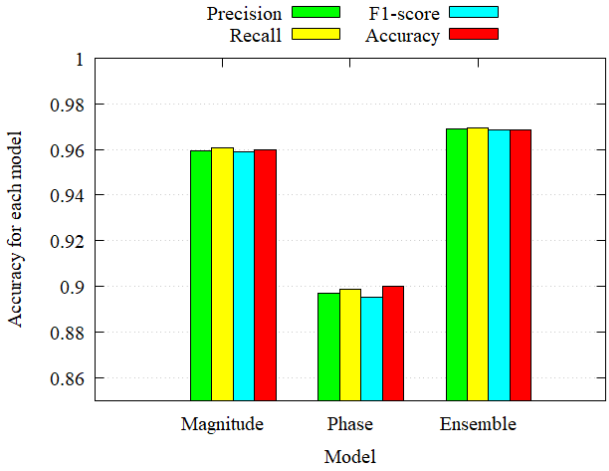

Figure 10: The performance comparison of each model.

We plotted the confusion matrices to also check the accuracy corresponding to each class. The confusion matrices are shown in figures 11– 13. Comparing the confusion matrices, it can be seen that the Mag and Phase models differ in the class that predicts well. For example, in Mag model, 'do nothing' gesture shows the highest accuracy, but not in Phase model. Since the accuracy of Mag model

is higher than that of Phase model, the ensemble model shows a similar pattern to Mag model. Also, almost classes of the ensemble model show better accuracy than Mag and Phase models.

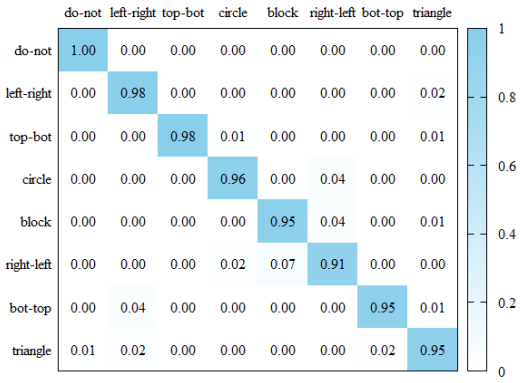

Figure 11: The confusion matrix of Mag model.

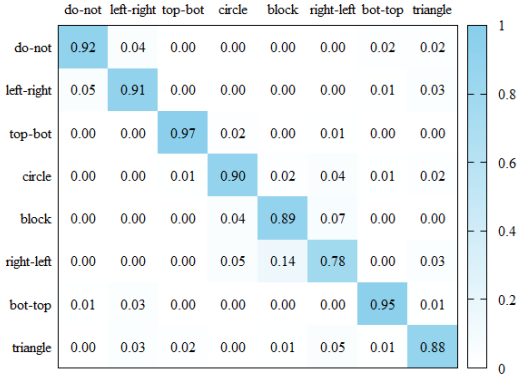

Figure 12: The confusion matrix of Phase model.

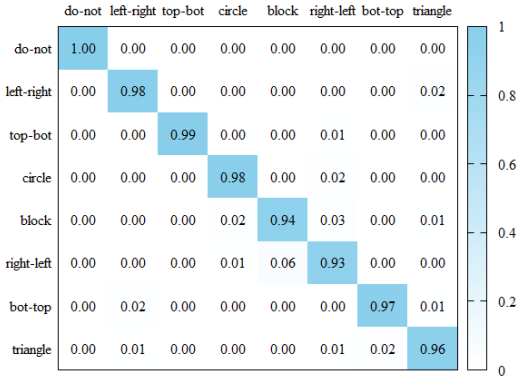

Figure 13: The confusion matrix of the ensemble model.

5 Conclusion

This study classified hand gestures using inaudible sound using a smartphone's speaker and microphone. We used our own application to produce and record 20 kHz sound while gesturing. The sound data were converted to STFT as magnitude and phase data. These data used for training and test in our CNN models. We proposed Mag and Phase models, and also suggested ensemble method using soft voting to improve classification accuracy. As a result, our model showed 96.88% classification accuracy for 8 hand gestures.

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

References

[1] G. Zhu, L. Zhang, P. Shen, J. Song, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," IEEE Access, **5**, 4517–4524, 2017, doi:10.1109/ACCESS.2017.2684186.

[2] W. Lu, Z. Tong, J. Chu, "Dynamic hand gesture recognition with Leap Motion controller," IEEE Signal Processing Letters, **23**(9), 1188–1192, 2016, doi:10.1109/LSP.2016.2590470.

[3] T. Schlömer, B. Poppinga, N. Henze, S. Boll, "Gesture recognition with a Wii controller," in Proceedings of the 2nd International Conference on Tangible and Embedded Interaction, 11–14, 2008, doi:10.1145/1347390.1347395.

[4] M. Sathiyanarayanan, S. Rajan, "MYO armband for physiotherapy healthcare: A case study using gesture recognition application," in 2016 8th International Conference on Communication Systems and Networks (COMSNETS), 1–6, 2016, doi:10.1109/COMSNETS.2016.7439933.

[5] R. Schafer, L. Rabiner, "Design and simulation of a speech analysis-synthesis system based on short-time Fourier analysis," IEEE Transactions on Audio and Electroacoustics, **21**(3), 165–174, 1973, doi:10.1109/TAU.1973.1162474.

[6] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," Nature, **521**(7553), 436–444, 2015, doi:10.1038/nature14539.

[7] H. Cheng, L. Yang, Z. Liu, "Survey on 3D hand gesture recognition," IEEE Transactions on Circuits and Systems for Video Technology, **26**(9), 1659–1673, 2016, doi:10.1109/TCSVT.2015.2469551.

[8] R. Nandakumar, V. Iyer, D. Tan, S. Gollakota, "FingerIO: Using active sonar for fine-grained finger tracking," in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 1515–1525, 2016, doi:10.1145/2858036.2858580.

[9] W. Wang, A. X. Liu, K. Sun, "Device-free gesture tracking using acoustic signals," in Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, 82–94, 2016, doi:10.1145/2973750.2973764.

[10] J. Kim, S. Choi, "Hand gesture classification based on nonaudible sound using convolutional neural network," Journal of Sensors, **2019**, 1084841, 2019, doi:10.1155/2019/1084841.

[11] T. G. Dietterich, "Ensemble methods in machine learning," in Multiple Classifier Systems, 1–15, 2000, doi:10.1007/3-540-45014-9_1.

[12] J. Schumacher, D. Sakič, A. Grumpe, G. A. Fink, C. Wöhler, "Active learning of ensemble classifiers for gesture recognition," in Pattern Recognition, 498–507, 2012, doi:10.1007/978-3-642-32717-9_50.

[13] Y. Lavinia, H. H. Vo, A. Verma, "Fusion based deep cnn for improved large-scale image action recognition," in 2016 IEEE International Symposium on Multimedia (ISM), 609–614, 2016, doi:10.1109/ISM.2016.0131.