논문 18-43-10-15 The Journal of Korean Institute of Communications and Information Sciences '18-10 Vol.43 No.10
https://doi.org/10.7840/kics.2018.43.10.1664

비가청 주파수를 이용한 손동작 분류

김 진 혁◆, 최 선 웅°

Hand Gesture Classification Using Inaudible Frequency

Jinhyuck Kim◆, Sunwoong Choi°

요 약

4차 산업혁명이 진행됨에 따라 인간의 행동이나 동작을 인식하고 구별하는 것이 중요한 이슈로 떠오르고 있다. 본 논문에서는 사람의 귀로 들을 수 없는 비가청 주파수 대역의 소리를 스마트폰에서 발생시키고, 그것이 반사된 신호를 녹음하여 손동작을 분류하는 방법을 제안한다. 제안하는 방법은 녹음된 소리 데이터를 단시간 푸리에 변환(Short-Time Fourier Transform)을 이용하여 이미지화 하고, Convolution Neural Network (CNN) 모델에 적용시켜 행동을 분류한다. 실험을 통해 제안하는 방법이 5개의 손동작에 대해서 94%의 정확도를 내는 것을 확인하였다.

Key Words : Inaudible Frequency, Gesture Classification, Convolution Neural Network, Short-Time Fourier Transform

ABSTRACT

As the 4th Industrial Revolution progresses, recognizing and distinguishing human actions and behaviors are becoming an important issue. In this paper, we propose a method to classify the hand gestures by generating the sound of inaudible frequency band that can not be heard by the human ear with the smartphone and recording the reflected signal. In the proposed method, the recorded sound data is imaged using a Short-Time Fourier Transform and applied to the Convolution Neural Network (CNN) model to classify the hand gestures. Experimental results show that the proposed method gives 94% accuracy for 5 hand gestures.

Ⅰ. 서 론

4차 산업혁명이 진행됨에 따라 인간과 컴퓨터간의 상호 작용(HCI : Human Computer Interaction) 기술이 중요해 지고 있다. 스마트 워치와 같은 각종 웨어러블 디바이스들이나 IoT 제품들이 늘어나면서 이를 쉽게 제어하기 위한 여러 가지 방법들에 대한 연구가 진행되고 있다. 광센서를 이용한 Okuli[1]나 RF 신호를 이용한 Google Soli[2]같은 연구들은 제스처 인식을 통해 특정 동작으로 디바이스를 제어하는 방식을 제안하고 있다. 하지만 두 가지 방법 모두 광센서나 RF 칩과 같은 추가적인 부품이 필요하다는 단점이 있다.

이에 따라, 추가적인 센서를 사용하지 않는 방법으로 음파를 사용하여 제스처를 인식하는 연구가 진행되었다. Microsoft Research에서 진행한 SoundWave[3] 연구에서는 이전 연구들과 다르게 별도의 변환기나 수신기를 사용하지 않고 노트북에 달려있는 상용 스피커와 마이크를 사용하였다. 노트북에 내장된 스피커

※ 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.2016R1A5A1012966)
◆ First Author : (ORCID:0000-0001-8727-5186)Kookmin University, Department of Secured Smart Electric Vehicle, j0910@kookmin.ac.kr, 정회원
° Corresponding Author : (ORCID:0000-0002-8719-8181)Kookmin University, School of Electrical Engineering, schoi@kookmin.ac.kr, 종신회원
논문번호 : 201807-221-D-RN, Received July 22, 2018; Revised September 10, 2018; Accepted September 18, 2018

와 마이크를 이용하여 연속적인 비가청 주파수를 발생시키고, 도플러 효과를 기반으로 하여 반사되어 들어오는 신호를 통해 제스처를 센싱하였다. 또 다른 제스처 인식 연구인 ER[4] 에서는 스마트폰을 이용하여 비가청 주파수를 발생시키고 도플러 효과를 기반으로 행동을 구분하였다. 위 연구에서는 SVM을 통해 차량 내부에서 하는 부주의한 행동 4가지를 분류하였으며, 94.8%의 정확도를 보였다.

본 논문에서는 별도의 센서 없이 비가청 주파수를 이용하여 손동작을 분류하는 것을 제안한다. 복잡한 필터 설계나 신호 처리 없이 도플러 효과를 기반으로 하여 반사된 신호의 특징을 추출한다. 반사되어 녹음된 신호를 이미지화 하고 딥 러닝 모델인 CNN(Convolution Neural Network)[5]에 적용시키는 방법을 통해 서로 다른 손동작을 구분하고 성능을 평가한다. 실험 결과 5가지 손동작에 대해서 94%의 분류 정확도를 보였다.

Ⅱ. 관련 연구

음파를 이용하여 손가락의 위치를 추적하거나 제스처를 분류하는 다양한 방법들이 연구되었다. 손가락의 위치를 추적하기 위한 연구 중 하나인 LLAP[6]는 음파의 위상변화를 이용한 연구로 Low-Latency Acoustic Phase (LLAP) 방식으로 위상 변화를 물체의 움직임의 길이로 변환하여 손가락 위치를 추적한다. 다른 연구인 FingerIO[7] 에서는 무선 통신에서 일반적으로 사용되는 변조 기술인 직교 주파수 분할 다중 방식(OFDM)을 이용한다. OFDM 방식을 통해 주파수 대역을 분할하고 각 주파수의 복소수의 실제 값을 변환을 통하여 거리를 계산한다. 마지막으로 Strata[8] 에서는 기존 논문들과 마찬가지로 음파를 이용하여 손가락 위치를 추적하는 연구를 진행하였다. 이 연구에서는 Channel Impulse Response (CIR)를 적용하여 반사되어 들어오는 다중 경로 신호 중 손가락에 해당하는 특정 채널을 추정한다. 추정한 채널의 위상변화를 기반으로 절대 거리와 상대 거리를 구하고 손가락 위치를 추적한다.

음파를 이용한 제스처 분류 연구 중 하나인 SoundWave[3]는 노트북을 이용하여 18~19kHz의 연속된 파일럿 톤의 소리를 출력하고, 반사되어 들어오는 신호를 분석하여 제스처를 분류한다. 총 5가지 제스처를 분류하였으며 96.6%의 분류 정확도를 보였다. 위 논문 같은 경우 별도의 머신러닝이나 딥러닝을 이용한 분류기를 사용하지 않고 단순히 반사된 신호의 특징을 수학적으로 계산하여 분류한 것이다. 또 다른 논문인 ER[4]은 음파를 이용하여 차량 내부에서 4가지 행동을 분류하였다. 차량 내부에 거치시킨 스마트폰을 이용하여 20kHz의 비가청 주파수에 해당하는 소리를 발생시키고 도플러 효과를 기반으로 각 행동별 특징을 추출한다. 이후 주성분분석(PCA) 기법과 SVM 모델을 이용하여 분류를 하였으며, 4가지 행동에 대하여 94.8%의 분류 정확도를 보였다.

본 논문에서는 기존에 음파를 이용하여 진행된 연구와 다르게 단시간 푸리에 변환(Short-Time Fourier Transform)[9]을 이용하여 시간에 대한 특정 주파수 대역의 데이터를 구하고, 딥 러닝 모델인 CNN (Convolution Neural Network)[5]을 이용한 손동작 분류 방법을 제안한다. 우리는 손가락의 위치를 정밀하게 추적하는 것이 아닌 단순한 특징 추출을 통해 손동작을 구분하는 것을 목표로 한다.

Ⅲ. 제안하는 방법

3.1 전체 시스템 구성

그림 1 은 제안하는 방법의 전체적인 시스템 구조이다. 먼저 우리는 직접 제작한 어플리케이션을 이용하여 녹음 데이터를 수집한다. 데이터 수집을 위하여 스마트폰 2대를 사용한다. 스마트폰 1대는 스피커 역할로 20kHz의 단일 대역 비가청 주파수를 일정 시간 동안 발생시킨다. 나머지 스마트폰 1대는 마이크 역할로 발생하는 신호를 녹음한다. 스마트폰이 녹음을 하는 동안 우리는 특정 행동을 수행하고, 이에 따라 각각 다른 반사 신호를 얻을 수 있게 된다.

다음으로 수집한 데이터를 PC로 옮기고 Matlab을 이용하여 STFT(Short-Time Fourier Transform)[6]를 적용시킨다. STFT를 통해 시간에 대한 주파수 대역별 세기로 데이터의 차원을 높혀 이미지화 한다. 이 과정에서 불필요한 주파수 대역은 버리고 비가청 주파수 대역만 잘라내서 사용한다. 잘라낸 데이터를 저장한다.

마지막으로 GPU 서버를 이용하여 CNN모델을 만

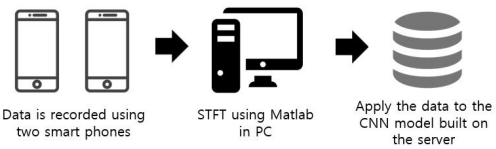


그림 1. 전체 시스템 구조
Fig. 1. Overall system architecture

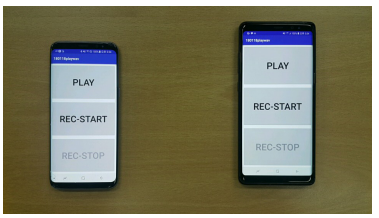
들고 데이터를 학습시켜 손동작 분류 성능을 평가한다. 우리는 Tensorflow를 사용하여 CNN 모델을 구현하였다. 저장된 데이터를 서버에서 받고, 학습 데이터와 테스트 데이터로 나누어 모델을 학습하고 성능을 평가한다.

3.2 실험 환경 및 구현

우리는 20kHz의 단일 비가청 주파수를 발생시키고 녹음하는 스마트폰 어플리케이션을 제작하였다. 그림 2 (a)는 제작한 어플리케이션의 UI이다. 어플리케이션은 안드로이드 스튜디오를 사용하여 구현하였으며, 특정 주파수 대역의 음파를 설정한 시간동안 발생시키고, 이를 녹음하는 기능을 수행한다. 녹음된 데이터는 해당 스마트폰 내부 저장소에 저장되는 방식이다. 본 논문에서는 20kHz의 비가청 주파수를 사용한다. PLAY 버튼을 누르면 10초 동안 20kHz의 비가청 주파수가 스마트폰 상단 스피커를 통해 재생된다. REC-START 버튼을 누르면 3초 동안 녹음을 진행한다. 녹음이 되는 3초 동안 REC-STOP 버튼이 활성화되며, 별도로 버튼을 누르지 않으면 녹음이 완료되고 자동으로 비활성화 된다.

실험을 위해 우리는 스마트폰 2대를 사용하였다. 비가청 주파수를 발생시키는 스피커 역할로 삼성 갤럭시 S8 모델을 사용하였다. 소리를 녹음하는 마이크 역할로는 삼성 갤럭시 노트8 모델을 사용하였다. 각각의 스마트폰에 제작한 어플리케이션을 설치하고 실험

(a)



(b)

그림 2. (a) 데이터 수집을 위해 제작한 어플리케이션 UI; (b) 실제 데이터 수집을 위한 실험 환경
Fig. 2. (a) Application UI created for data collection; (b) Experimental environment for actual data collection

을 진행한다.

실험은 사람이 없는 실험실에서 진행하였다. 테이블 위에 스마트폰 2대를 간격을 두고 올려놓은 다음 스피커 역할에 해당되는 스마트폰에서 PLAY버튼을 누른다. 다음으로 마이크 역할에 해당되는 스마트폰에서 REC-START버튼을 누르고 특정 손동작을 취한다. 위와 같은 행동을 반복하면서 데이터를 수집하였다. 그림 2 (b)는 실제 데이터 수집을 위한 실험 환경의 모습이다.

3.3 제안하는 알고리즘

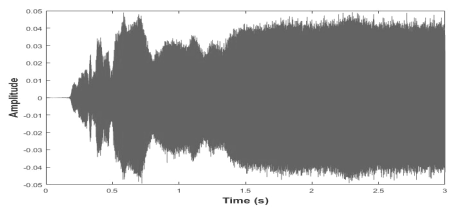
우리는 어플리케이션을 이용하여 수집한 데이터에 Matlab을 이용하여 STFT를 적용시킨다. 이를 통해 시간에 따른 주파수 대역별 데이터를 얻을 수 있게 된다. 이후 관심 영역인 비가청 주파수 대역에 해당되는 19.8kHz~20.2kHz구간만 잘라내어 사용한다.

데이터는 44.1kHz의 샘플링 주파수로 3초 동안 녹음 된다. 녹음된 데이터에서 녹음 시작 시 발생하는 시스템 딜레이를 제거하기 위해 시작 0.2초를 잘라내어 사용한다. 따라서 STFT에 사용되는 데이터는 44.1kHz의 샘플링 주파수를 가진 2.8초짜리 데이터이다.

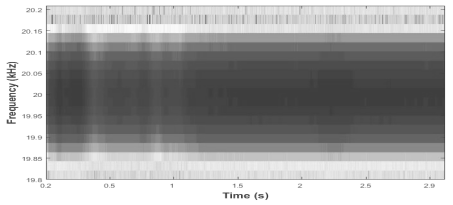
다음으로 시간에 따른 특정 주파수 대역의 값을 구하기 위해 데이터에 STFT를 적용시킨다. STFT를 적용시킬 때 주파수 Resolution은 2048, Window Size를 500으로 설정하고 95%씩 오버랩 시킨다. 샘플링 주파수가 44.1kHz인 신호이기 때문에, Resolution을 2048로 설정하면 주파수 한 구간 당 약 21Hz를 나타나게 된다. 따라서 관심 주파수 대역인 19.8kHz~20.2kHz 구간은 20개의 주파수 구간으로 나누어진다. 그리고 44.1kHz 샘플링 주파수를 갖는 2.8초짜리 데이터에 Window Size를 500으로 95%씩 오버랩 시키면 4920개의 시간 구간으로 나누어진다. 결과적으로 STFT 이후 얻어지는 데이터의 사이즈는 20x4920x1 이다.

그림 3 (a) 는 마이크를 막고 녹음한 데이터를 시간에 대한 음파의 크기의 그래프로 나타낸 것이다. 그림 3 (b) 는 동일한 데이터에 STFT를 적용시킨 후 비가청 주파수 대역을 잘라내어 시간에 대한 주파수 영역의 세기로 나타낸 것이다. 특정 주파수에서 신호의 세기가 강할수록 암적색을 신호의 세기가 약할수록 푸른색을 띄고 있다.

다음으로 STFT 결과로 얻어진 데이터를 CNN 모델에 학습시킨다. 그림 4는 우리가 제안하는 CNN 알고리즘이다. 제안하는 CNN 알고리즘은 9층으로 구성된 모델이다. 구조는 입력 데이터의 사이즈를 고려하



(a)



(b)

그림 3. (a) 마이크를 막았을 때 녹음된 파형; (b) STFT를 적용시킨 후 비가청 주파수 대역을 잘라낸 결과
Fig. 3. (a) Waveform of recorded sound wave when microphone is blocked; (b) The result of cutting out the non-audible frequency band after applying STFT

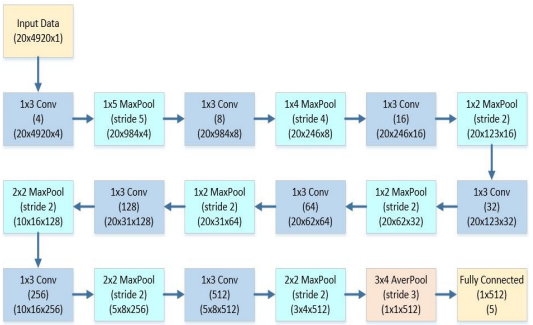


그림 4. 제안하는 CNN 알고리즘
Fig. 4. Our proposed CNN algorithm

여 Filter Size를 정하고 Convolution과 Max Pooling을 반복하도록 구성하였다. 다음으로 Fully Connected를 하기 전 한번의 Average Pooling 을 통해 해당 영역의 평균을 구하여 데이터의 크기를 줄인다. 마지막으로 Fully Connected를 하여 원하는 Label 수만큼의 출력을 얻는다.

Ⅳ. 실 험

4.1 실험 데이터

실험에서는 총 5가지 종류의 손동작 데이터를 수집하고 사용하였다. 분류하는 5가지 손동작은 아래와 같다.

(1) Do nothing : 녹음이 진행되는 동안 아무런 행동을 하지 않는다.

(2) Move from left to right : 녹음이 진행되는 동안 손바닥을 편 상태에서 손을 왼쪽에서 오른쪽으로 움직인다. 이때 손은 녹음을 하는 스마트폰 액정 위에 위치한다.

(3) Move from top to bottom : 녹음이 진행되는 동안 손바닥을 편 상태에서 손을 위쪽에서 아래쪽으로 움직인다. 이때 손은 녹음을 하는 스마트폰 액정 위에 위치한다.

(4) Circle drawing : 녹음이 진행되는 동안 검지 손가락을 편 상태에서 시계방향으로 원을 그린다. 이때 손가락은 녹음을 하는 스마트폰 액정 위에 위치한다.

(5) Block the microphone : 녹음이 진행되는 동안 스마트폰 하단에 있는 마이크 부분을 손바닥을 이용하여 막는다.

각 데이터는 3초씩 녹음하였으며, 한 종류의 손동작 당 100번씩 총 500개의 데이터를 수집하였다. 각 데이터의 Sample rate는 44.1kHz이다. 시스템 딜레이 0.2초를 잘라내었기 때문에 실제 사용하는 데이터는 2.8초이다. 표 1 은 실험에 사용한 데이터를 표로 나타낸 것이다.

표 1. 사용한 녹음 데이터
Table 1. Used recording data

Gesture	Number of data
Do nothing	100
Left to right	100
Top to bottom	100
Drawing circle	100
Block the microphone	100

4.2 실험 결과

STFT 한 후 데이터를 CNN 모델에 학습시키고 성능을 평가하였다. 5가지 손동작에 대하여 각각 100개씩 총 500개의 데이터를 사용하였다. 데이터를 8:2로 나누어 총 500개의 데이터 중 400개를 모델 훈련에 사용하였고, 100개의 데이터로 테스트를 진행하였다.

테스트 결과 전체 정확도는 94%라는 결과를 보였다. 그림 5는 평가 결과를 Confusion matrix로 나타낸 것이다. 각 행동별로 Do nothing 과 Move from top to bottom 행동에 대해서는 100%의 분류 정확도를 보였고, Move from left to right 행동은 95%의 분류 정확도로 5%를 Move from top to bottom으로 맞지 않게 분류하였다. Circle drawing 행동에 대해서는 90%의 분류 정확도를 보였으며 각각 5%씩 Move

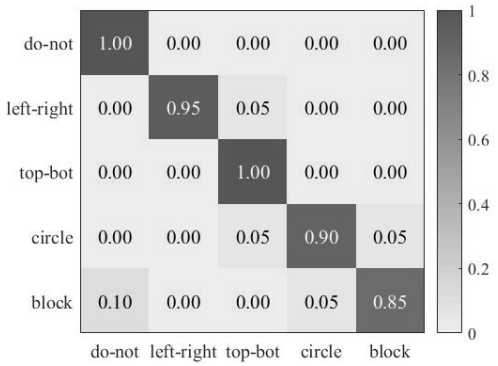


그림 5. STFT한 데이터를 CNN 모델에 적용한 결과 Confusion matrix
Fig. 5. The result of applying the STFT data to the CNN model show the confusion matrix

from top to bottom 과 Block the microphone 행동으로 분류하였다. 마지막으로 Block the microphone 행동에 대해서는 Do nothing으로 10%, Circle drawing으로 5% 맞지 않게 분류하여 85%의 정확도를 보였다. 결과적으로 각 행동별 85% 이상의 분류 정확도를 보였다.

추가적으로 우리는 CNN이 아닌 데이터 분류에 사용되는 다른 기계 학습을 사용하여 결과를 비교하였다. 입력 데이터로는 STFT를 적용한 후의 데이터를 사용하였다. 비교에 사용한 기계 학습은 Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF) 총 3가지이다. 서버에 Python으로 Scikit-learn 사용하여 각각의 알고리즘을 구현하고 성능을 평가하였다. 그림 6 은 3가지 기계 학습 알고리즘과 제안한 방법의 결과 정확도를 비교하여 나타낸 것이다. 평가 결과 정확도는 순서대로 57%, 64%, 78.7%의 결과를 보였다. 1차원 데이터로 녹음된 신호를 STFT를 이용하여 2차원 데이터로 변환하고, 이미지 분류에 높은 정확도를 보이는 CNN 모델을 이용하여 1차원 데이터로부터 얻을 수 없는 특징을 구하고 이를 통해 분류 정확도를 높였다. 결과적으로 제안한 방법이 기존에 분류 알고리즘에 비해 높은 분류 정확도를 보였다.

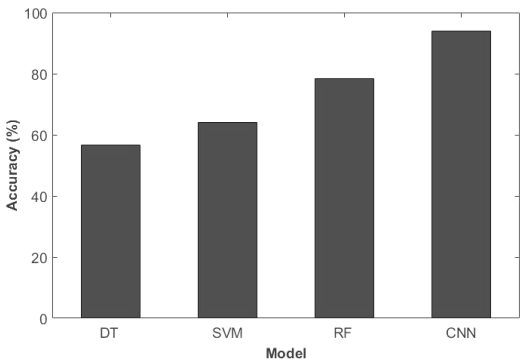


그림 6. 다른 분류 알고리즘과 제안한 방법과의 정확도 비교
Fig. 6. Comparison of Accuracy between another classification algorithms and proposed method

V. 결 론

본 논문에서는 스마트폰 비가청 주파수를 이용하여 손동작을 분류하였다. 데이터 분류를 위해 STFT를 이용하여 시간에 따른 주파수 특성을 구하고 CNN 모델에 적용시키는 방법을 제안하였다. 제안한 방법은 5가지 행동에 대해 94%의 분류 정확도를 보였으며, 각 행동별로 85%이상의 결과를 보였다. 또한 다른 기계 학습 모델 3가지와 비교하였을 때 더 높은 분류 정확도를 보였다.

향후 다른 동작들을 추가하고 장소 및 환경을 변경해 보면서 추가 실험을 진행 할 예정이다. 또한 STFT를 하지 않은 Raw 데이터를 사용하여 학습하는 CNN 모델을 구축하고 결과를 비교해 볼 것이다. 추가적으로 학습 모델을 만들고 어플리케이션과 연동하여 실시간으로 평가 가능한 시스템을 구축한다면, 제스쳐 인식을 이용한 다양한 연구가 가능할 것으로 기대된다.

References

[1] C. Zhang, J. Tabor, J. Zhang, and X. Zhang, "Extending mobile interaction through near-field visible light sensing," *MobiCom*, pp. 345-357, Sep. 2015.

[2] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph*, vol. 35, no. 4, Article 142, Jul. 2016.

[3] S. Gupta, D. Morris, S. N. Patel, and D. Tan, "SoundWave: Using the doppler effect to sense gestures," *CHI*, pp. 1911-1914, May 2012.

[4] H. Gao, X. Xu, J. Yu, Y. Chen, Y. Zhu, G. Xue, and M. Li, "ER: Early recognition of inattentive driving leveraging audio devices on

smartphones," *INFOCOM*, pp. 1-9, May 2017.
[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
[6] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," *MobiCom*, pp. 82-94, Oct. 2016.
[7] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using active sonar for fine-grained finger tracking," *CHI*, pp. 1515-1525, May 2016.
[8] S. Yun, Y. C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," *MobiSys*, pp. 15-28, Jun. 2017.
[9] J. B. Allen, "Short term spectral analysis, synthesis, and, modification by discrete fourier transform," in *Proc. IEEE Acoustic, Speech, and Sign. Process.*, vol. 25, no. 3, pp. 235-238, Jun. 1997.

김 진 혁 (Jinhyuck Kim)

2017년 2월 : 국민대학교 전자공학부 졸업
2017년 3월~현재 : 국민대학교 보안-스마트 전기자동차학과 석사과정
<관심분야> 기계학습, 사물인터넷, 임베디드 시스템

최 선 웅 (Sunwoong Choi)

1998년 2월 : 서울대학교 전산과학과 졸업
2000년 2월 : 서울대학교 전산과학과 석사
2005년 8월 : 서울대학교 전기, 컴퓨터공학부 박사
2007년 3월~현재 : 국민대학교 전자공학부 부교수
<관심분야> 유무선 네트워크, 기계학습, 사물인터넷