소리 정보를 이용한 딥 러닝 기반의 동작 인식

차주형[1] · 배성준[1] · 박지은[1] · 이준혁[1] · 장시웅[1] · 이현섭[1]

[1]동의대학교

Deep Learning based Motion Recognition Using Sound Event

Joo Hyoung Cha[1] · Seong Jun Bae[1] · Jieun Park[1]

Jun Hyeok Lee[1] · Si-Woong Jang[1] · HyunSup Lee[1]

[1]Dong Eui University

E-mail : chacha@udon.party

요 약

설치된 마이크를 통해 사용자의 행동에 대한 음성 데이터를 수집하고, fast fourier transform(FFT)와 딥 러닝을 통해 동작되는 행동에 대해 식별이 가능한지 검증한다. 본 실험의 데이터는 3개의 경우의 수로 각각 걷는 동작, 문을 여는 동작, 닫는 동작으로 정의하였고, 내부 실험을 위해 상황마다 약 200개의 데이터를 수집하였다. 본 논문에서는 수집된 데이터와 딥 러닝의 기본적인 모델 CNN을 통해 98%의 정확도를 얻었으며, 음성 데이터를 활용하여 사용자의 행동을 충분히 식별을 할 수 있음을 보여준다. 향후, 음성의 세기에 따른 삼각 측량을 수행하여 행동뿐만 아니라 행위자의 위치 예측 모델을 수행할 예정이다.

키워드

fast fourier transform, prediction, neural network, audio analysis

Ⅰ. 서 론

최근에 시각적인 정보를 활용하여 영상 데이터 가공에 관한 연구가 많이 진행되고 있다. 공간과 특징 추출에 특화된 Convolution 기반의 딥 러닝 모델은 특징 추출에 효과적이다. 이를 활용하여 공장 자동화, 자율 주행 자동차, 방범 시스템과 같은 분야에서 주로 시각적인 정보를 활용하여 수행되고 있다[1][2].

시각적인 정보 기반의 데이터를 활용한 인공지능은 제한된 공간과 특정 시나리오에서는 정상적으로 동작하지만, 악천후나 의도적인 카메라 가림과 같은 변인 통제가 어려운 상황에서는 높은 신뢰성을 기대하기 어렵다. 이에, 시각적인 정보와 더불어 다른 데이터를 함께 활용하거나 그 외의 데이터 처리가 필요하다[3].

본 논문에서는 행동을 식별하기 위해 카메라와 같은 영상 정보가 아닌, 사용자의 움직임으로 인한 진동 계수를 활용한다. 시간-도메인 영역을 주파수-도메인 영역으로 변환하여 행동에 따른 고유 주파수를 분석하고, 이를 딥 러닝에 학습한 결과에 따른 신뢰성 분석을 수행한다[4].

본 논문의 구성은 다음과 같다. Ⅱ장에서 실험 환경과 관련 연구에 대해 설명하고, Ⅲ장에서 수집된 데이터를 기반으로 실험 결과를 분석한다. 마지막으로 Ⅳ장에서 실험에 대한 결론을 맺는다.

Ⅱ. 실험 정의

음성 정보를 활용하여 사용자 행동을 식별하기 위해, 우선 행동의 종류에 대해 정의가 필요하였고, 크게 3개의 행위로 정의하였다. 첫 번째는 보행자의 움직임과 같은 이동이며, 두 번째와 세 번째는 독립된 방안을 입실과 퇴실하는 것으로 정의하였다.

제한된 환경이 아닌, 일상 환경에서 실험하기 위해 잡음(회의, 잡담, 노래)을 포함하였고, 잡음 정보의 필터링 과정은 진행하지 않았다.

잡음 데이터와 필요 정보가 섞인 시간 도메인 정보를 유의미한 정보로 변환하기 위해 fourier transform을 수행하여 데이터 전처리 과정을 수행

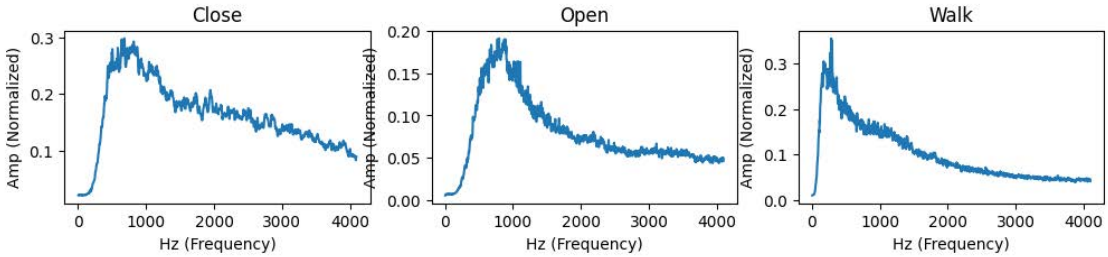

Fig. 1. Arithmetic mean value of normalized frequency domain

하였다.

임의의 입력되는 신호에서 다양한 주파수를 갖는 주기 함수들의 합으로 분해하여 표기하는 주파수 도메인으로 변환하는 과정을 수행하기 위해서 시간과 Hz, 세기 정보가 필요하다. 이후, 변환된 주파수의 정보를 활용하여, 랜덤한 임의의 주파수를 가지는 잡음 정보에 대한 필터링이나 원하는 정보 추출이 가능하다.

그림 1은 각각의 행동에 따른, 정규화된 주파수 도메인의 평균 그래프이다. 퇴실과 입실, 보행 모두 0부터 1,000Hz의 범위의 주파수에서 신호가 발생하는 것을 알 수 있다. 또한, 퇴실과 입실의 경우, 모두 문의 움직임 인한 소리가 발생하면서 서로 주파수 대역이 유사한 모습을 나타낸다.

Table. 1. Audio data set for training and

loc. / type	Open	Close	Walk
times	205	206	227
avg time	4 ~ 6 s	4 ~ 6 s	0 ~ 1 s
total time	10m 15s	10m 18s	1m 53.5s

데이터의 셋의 경우 데이터 유형마다 녹음한 음성 데이터의 수를 나타낸다. 원시 음성 데이터의 경우 일정한 주기를 가지는 데이터가 아니므로, 수기로 데이터 유형마다 약 200개의 데이터를 수집하였다. 음성 파일마다 약 4초의 정보를 담고 있으며, 동의대학교 산학협력관의 4층에서 녹음한 소리이다.

수집된 데이터를 FFT 수행한 주파수 도메인의 값인 1부터 4,096Hz의 범위의 값을 H(64), W(64), C(1)의 입력 데이터셋으로 변환하여 학습을 수행하였다. 딥 러닝 모델은 CNN을 사용하여 성능을 검증한다[5].

III. 실험 결과

본 연구에서 3개의 상황에 대해 638개의 음성 데이터를 생성하였고, FFT 연산을 수행하여 주파수 도메인의 정보를 학습 정보로 활용하였다.

638개의 데이터 중, 랜덤으로 510개와 128개로 분리하여 학습 데이터와 검증 데이터 셋으로 나눠 수행하였다[6]. 학습 결과, 4 에폭에서 97%의 정확도가 나타났으며, 100 에폭까지 학습한 결과 최대 98%의 정확도를 나타내었다.

IV. 결론

본 논문에서는 시간 속성을 제외하고, 현재 상황에 대해 분석하여 행동 식별이 가능한지 확인하였다. 향후, 시간 속성과 이종 데이터를 포함하여 강인한 동작 인식 알고리즘에 대해 연구를 수행할 예정이다.

또한, 1개의 음성 녹음기가 아닌, 복수의 음성 녹음기를 통해, 위상 스펙트럼 분석을 수행하여 행위자의 위치와 행동 추론 알고리즘에 대해 연구를 수행할 예정이다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성(Grand ICT연구센터) 사업의 연구결과로 수행되었음 (IITP-2023-2020-0-01791)

본 연구는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 3단계 산학연협력 선도대학 육성사업(LINC 3.0)의연구결과입니다.

References

[1] [1] Kim P, Huang X, Fang Z. SSD PCB Component Detection Using YOLOv5 Model. J. Inf. Commun. Converg. Eng. 2023;21:24-31. https://doi.org/10.56977/jicce.2023.21.1.24

[2] Ahn Y, Kim KB, Park H. Comparison of the Effect of Interpolation on the Mask R-CNN Model. J. Inf. Commun. Converg. Eng. 2023;21:17-23. https://doi.org/10.56977/jicce.2023.21.1.17

[3] Su-yeon Han and Dea-Woo Park. "Cat Behavior

Pattern Analysis and Disease Prediction System of Home CCTV Images using AI." Journal of t he Korea Institute of Information and Communi cation Engineering, vol. 26, no. 9, 2022, 1266-1 271.

[4] T. Kobayashi, A. Kubota and Y. Suzuki, "Audi o Feature Extraction Based on Sub-Band Signal Correlations for Music Genre Classification," 20 18 IEEE International Symposium on Multimedi a (ISM), Taichung, Taiwan, 2018, pp. 180-181, doi: 10.1109/ISM.2018.00-15.

[5] Eom Y, Bang J. Speech Emotion Recognition Using 2D-CNN with Mel-Frequency Cepstrum C oefficients. J. Inf. Commun. Converg. Eng. 202 1;19:148-154. https://doi.org/10.6109/jicce.2021.1 9.3.148

[6] Paszke, A., Gross, S., Massa, F., Lerer, A., Bra dbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural inform ation processing systems, 32.