Early Fusion을 적용한 위급상황 음향 분류

양진환[1], 김성식[1], 최혁순[1], 문남미[2]
[1]호서대학교 컴퓨터공학부 학부생
[2]호서대학교 컴퓨터공학부 교수
yjh970706@naver.com, kim07084@naver.com
hucksoon2001@gmail.com, nammee.moon@gmail.com

Emergency Sound Classification with Early Fusion

Jin-Hwan Yang[1], Sung-Sik Kim[1], Hyuk-Soon Choi[1], Nammee Moon[1]
[1]Dept. of Computer Engineering, Hoseo University

요 약

현재 국내외 CCTV 구축량 증가로 사생활 침해와 높은 설치 비용등이 문제점으로 제기되고 있다. 따라서 본 연구는 Early Fusion을 적용한 위급상황 음향 분류 모델을 제안한다. 음향 데이터에 STFT(Short Time Fourier Transform), Spectrogram, Mel-Spectrogram을 적용해 특징 벡터를 추출하고 3차원으로 Early Fusion하여 ResNet, DenseNet, EfficientNetV2으로 학습한다. 실험 결과 Early Fusion 방법이 가장 좋은 결과를 보였고 DenseNet, EfficientNetV2가 Accuracy, F1-Score 모두 0.972의 성능을 보였다.

1. 서론

현재 국내외 CCTV 구축량은 증가하고 있는 추세이다[1]. 그로인한 범죄 예방, 증거 활용 등의 장점에도 불구하고 사생활 침해와 높은 설치 비용 등이 문제점으로 제기되고 있다[1]. 또한 ICT 기술의 발전으로 CCTV가 점차 고도화, 지능화됨에 따라 설치에 대한 부정적인 인식을 줄 가능성이 있다[2]. 하지만 CCTV 설치 시 범죄 억제 효과를 무시할 수 없으므로 새로운 대안이 필요하다[1,2].

따라서 본 연구는 CCTV의 사생활 침해 방지 및 설치 비용 절감을 위해 영상 데이터 없이 음향 데이터만으로 위급상황을 감지할 수 있는 인공지능 위급상황 음향 분류 모델을 제작한다.

음향 분류 모델 학습을 위해 STFT, Spectrogram, Mel-Spectrogram을 활용해 특징 벡터를 추출하고 3차원으로 Early Fusion하여 모델 학습에 활용하는 방법을 제안한다[3].

Early Fusion 방법의 실증을 위해 각 특징 벡터 추출 방법과 Early Fusion 방법의 학습 결과를 비교한다. 그 후 가장 좋은 결과를 보인 방법으로 이미지 분류 모델인 ResNet, DenseNet, EfficientNetV2를 이용해 학습하고 결과를 비교한다[4,5,6].

2. 데이터셋 전처리

2.1 위급상황 음성/음향 데이터셋

학습에 활용한 데이터셋은 AIHub의 “위급상황 음성/음향”으로 데이터셋의 규모는 총 402,767개이며 각 클래스별 분포는 <표 1>과 같다.

<표 1> 데이터셋 분포표

분류	규모(단위:개)	분류	규모(단위:개)
강제추행	20,945	가스사고	17,603
강도범죄	19,490	붕괴사고	29,691
절도범죄	12,646	태풍-강풍	23,350
폭력범죄	21,183	지진	28,240
화재	47,401	도움요청	40,734
갇힘	13,029	실내	30,963
응급의료	46,897	실외	12,085
전기사고	38,530	계	402,767

2.2 STFT

STFT는 음향 신호를 짧은 구간으로 나누어 푸리에 변환을 적용한다. 일반 푸리에 변환이 데이터의 시간적 변화를 담지 못하는 것에 비해 STFT는 시간적 변화와 주파수적 변화를 모두 파악할 수 있다.

2.3 Spectrogram

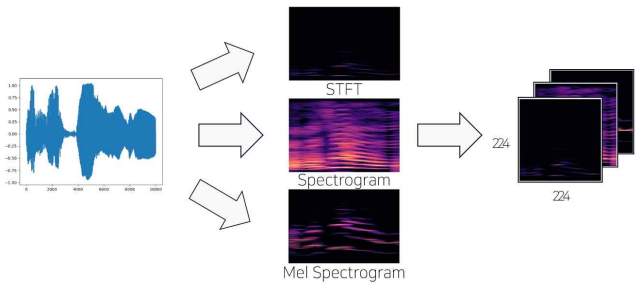
음향 신호에 STFT를 적용한 후 Magnitude 성분에 dB 스케일을 취하면 Spectrogram을 얻을 수 있다. dB 스케일은 인간이 저음역대 변화에 더 민감한 신체적 특징을 반영한 로그 스케일을 뜻한다.

2.4 Mel-Spectrogram

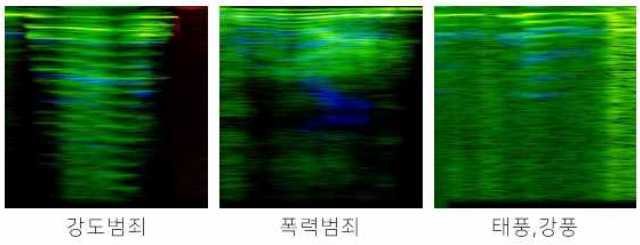
음향 신호에 STFT를 적용한 후 Mel Filter Bank를 적용하면 Mel-Spectrogram을 얻을 수 있다. Mel Filter Bank는 인간의 귀가 저음역 변화에 더 민감한 것을 반영한 필터뱅크이다.

2.5 Early Fusion

본 연구는 하나의 데이터에 STFT, Spectrogram, Mel-Spectrogram을 이용해 세 가지 특징 벡터를 추출하고 이 특징 벡터들을 각각 이미지의 R, G, B 채널로 치환하여 Early Fusion한다. Early Fusion의 개념도는 (그림 1), 제작된 이미지 예시는 (그림 2)와 같다.



(그림 1) Early Fusion 개념도



(그림 2) Early Fusion 이미지 예시

3. 실험

3.1 실험 세부사항

제안한 Early Fusion 방법의 실증을 위해 STFT, Spectrogram, Mel-Spectrogram특징 벡터로 학습한 결과와 Early Fusion으로 학습한 결과를 비교한다.

실험은 전체 데이터에서 클래스별로 500개씩, 총 7,500개를 추려서 6:2:2 비율로 Train, Validation, Test 데이터셋을 나누고 이미지넷으로 사전 학습된 Resnet50 모델로 학습하여 결과를 비교한다.

그 후 가장 좋은 결과를 보이는 방법으로 총 402,767개의 데이터에서 특징 벡터를 추출하고 6:2:2 비율로 Train, Validation, Test 데이터셋을 나누어 이미지넷으로 사전학습된 ResNet50, DenseNet121, EfficientNetV2s 모델로 학습하고 결과를 비교한다.

3.2 실험 결과

특징 벡터 추출 방법 실험 결과는 <표 2>와 같다.

<표 2> 특징 벡터 추출 방법 실험결과표

Feature	Accuracy	F1-Score
STFT	0.732	0.707
Spectrogram	0.744	0.717
Mel-Spectrogram	0.734	0.709
Early Fusion	**0.755**	**0.727**

실험 결과 Early Fusion 방법이 각 특징 벡터 추출 방법을 따로 학습한 것보다 좋은 성능을 보였다.

Early Fusion 방법을 활용해 이미지 모델을 학습한 결과는 <표 3>과 같다.

<표 3> 모델 비교 실험결과표

Model	Accuracy	F1-Score
ResNet50	0.971	0.971
DenseNet121	0.972	0.972
EfficientNetV2s	0.972	0.972

4. 결론

본 연구에서는 음향 분석에 쓰이는 푸리에 변환 기반 특징 벡터 추출 방식인 STFT, Spectrogram, Mel-Spectrogram을 이용해 추출한 특징 벡터에 Early Fusion을 적용하여 위급상황 음향을 분류하는 모델을 제작하였다.

Early Fusion 방식은 각 특징 벡터를 따로 학습한 것보다 더 좋은 정확도를 보였으며 DenseNe121t과 EfficientNetV2s가 Accuracy와 F1-Score 모두 0.972의 성능을 보였다.

ACKNOWLEDGEMENT

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF- 2021R1A2C2011966).

참고문헌

[1] 김익회, 이재용, “공공 방범 CCTV 의 국내 확산을 위한 방안 연구”, 한국지리학회지, 8(1), pp. 79-93, 2019

[2] 이주영, 조윤오, “범죄 예방을 위한 공공 CCTV 증설 인식 연구: 보호동기이론의 적용을 중심으로”, 한국범죄학, 제17권, 제1호, pp. 45-63., 2023

[3] JeongHyeon Park, JunHyeok Go, SiUng Kim, Nammee Moon, “Classification of Infant Crying Audio based on 3D Feature-Vector through Audio Data Augmentation“, 한국컴퓨터정보학회논문지, Vol. 28, No. 9, pp. 47-54, September 2023

[4] He, K., Zhang, X., Ren, S., & Sun, J., “Deep residual learning for image recognition”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016

[5] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q., “Densely connected convolutional networks”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708, 2017

[6] Tan, M., & Le, Q., “Efficientnetv2: Smaller models and faster training”, In International conference on machine learning, pp. 10096-10106, 2021