멀티모달 기반 지능형 위험 탐지 시스템

이서아, 홍용근, 정관수*, 노학균

대전대학교, *교원대학교

seoasap907@naver.com, yghong@dju.kr, *ksjung@knue.ac.kr, hg.roh@dju.kr

Design and Evaluation of a Multimodal Intelligent Risk Detection System

Seoa Lee, Yong-Geun Hong, Kwansoo Jung*, Hakgyun Roh

Daejeon University, *Korea National Univ. of Education

요 약

본 논문은 1인 가구 및 고령자 등 사회적 약자의 주거 안전을 위협하는 현관 범죄에 대응하기 위해, 멀티모달 기반의 지능형 위험 탐지 기법(Intelligent Risk Detection System)을 제안한다. 기존의 단일 센서 방식이 갖는 탐지 사각지대를 보완하고자, 객체 인식과 음성 인식을 결합하였다. 객체 인식을 위해서 YOLOv8 모델을 최적화하여 위험 객체(칼, 방망이 등)의 탐지 정밀도를 높였으며, 음성 인식을 위해서는 STT(Speech-to-Text) 및 SVM 분류기를 통해 비명이나 위협성 발화를 실시간으로 식별하였다. 또한, 서로 다른 모달리티의 탐지 결과에 가중치를 부여하여 멀티모달 기반의 위험 판단의 신뢰성을 강화하였다. 구현된 프로토타입은 복합적인 위협 상황을 효과적으로 인지하고 즉각적인 경보를 제공함으로써, 주거 취약 계층을 위한 능동적 위험 감지 솔루션으로서의 가능성을 입증하였다.

Ⅰ. 서 론

1인 가구 증가와 고령화로 인해 현관 침입·위협 등 주거 범죄 위험이 커지면서, 사회적 약자의 안전을 강화할 수 있는 기술적 대응이 요구된다. 그러나 기존의 단일 센서 기반 탐지는 한계가 뚜렷하다. 예를 들어 이미지 기반은 조도 저하·역광·가림(사각지대) 상황에서 위험 물체를 놓치기 쉽고, 음성 기반은 생활소음이 섞인 환경에서 비명·협박 등 위험 신호를 안정적으로 분리하기 어렵다.

본 연구는 이러한 문제를 해결하기 위해 스마트 초인종 환경을 위한 지능형 위험 탐지 시스템을 제안한다. 제안 시스템은 Vision(객체 인식)과 Voice(음성 인식)를 결합한 멀티모달 구조로, YOLOv8을 통해 칼·방망이 등 위험 물체를 검출하고 SVM으로 비명·협박 등 위험 소리를 분류한다. 이를 통해 위험 물체와 비위험 물체를 보다 정확히 구분하고, 소음 혼재 상황에서도 위험 여부 판단의 신뢰성을 높인다. 또한 위험 상황 발생 시 사용자에게 즉각 경고를 제공하여 실시간 대응 가능성과 안전성을 향상시킨다.

1. 관련 연구

스마트 초인종 설계에 연합학습 기반 딥러닝을 적용한 연구[1]는 주거 환경에서의 지능형 감지 시스템 구현 가능성을 제시하며, 실제 서비스 환경에서도 지속적으로 모델을 개선할 수 있는 방향을 보여준다. 약지도 학습 기반의 멀티모달 폭력 탐지 연구[2]는 영상(위험 행동/상황)과 음성(비명·협박 등 위험 발화) 정보를 함께 학습해, 단일 모달로 놓치기 쉬운 위험 신호를 더 안정적으로 포착할 수 있음을 보고하였다.

또한 오디오-비주얼 이벤트 인식을 위한 주의 기반 융합 네트워크 연구[3]는 음성·영상 특징을 결합하는 융합 전략이 복합 환경에서 사건 인식 성능을 높이는 데 유효함을 보여주며, 위험 객체 및 위험 발화 탐지와 같은 멀티모달 위험 인식 문제에 적용 가능함을 보여준다.

Ⅱ. 본론

본 절에서는 이미지 기반 객체 인식 모듈과 음성 기반 위험 단어 탐지 모듈을 결합한 멀티모달 기반 지능형 위험 탐지 기법을 설명한다.

A. 객체 인식 설계 및 학습 과정

YOLOv8 모델의 크기별 성능(n, s, m)을 비교하여 탐지 정확도 향상을 위한 최적 모델을 선정하였다. 제안된 이미지 기반 객체 인식 모듈은 방문자 이미지로부터 위험 물체를 탐지하며, 칼·방망이와 같은 위험 객체와 우산·나무스틱과 같은 비위험 객체를 분류하도록 구성하였다.

1) 위험 물체와 안전 물체의 분류

모델 학습은 칼과 방망이를 위험 물체, 우산과 나무스틱을 비위험 물체로 분류하여 진행하였다. 형태적 유사성을 갖는 객체를 포함하도록 데이터셋을 구성함으로써, 모델의 정밀한 시각적 인식 능력과 분류 성능을 검증하였다.

2) 객체 인식(Object Detection)

칼, 방망이, 우산, 나무스틱에 대한 이미지 데이터셋은 Roboflow를 통해 수집하였으며, 이후 라벨링을 수행한 뒤 전처리 과정을 거쳐 학습에 활용하였다. 객체 인식 모델로는 YOLOv8을 적용하여 물체 탐지 학습을 진행하였다.

초기 실험에서는 경량화된 YOLOv8n 모델을 적용하여 탐지 정확도가 낮게 나타났으나, 모델의 표현 능력을 강화하기 위해 YOLOv8s 및 YOLOv8m 모델을 순차적으로 학습하여 성능을 개선하였다.

Table 1. YOLOv8n 모델 성능

Class	Precision	Recall	mAP50
Umbrella	0.12	0.707	0.212
Knife	0.934	0.118	0.228
Stick	0.627	0.713	0.737
Bat	0.712	0.551	0.63
All (Mean)	**0.645**	**0.495**	**0.454**

Table 2. YOLOv8m 모델 성능

Class	Precision	Recall	mAP50
Umbrella	0.785	0.707	0.806
Knife	0.760	0.667	0.736
Stick	0.710	0.775	0.813
Bat	0.736	0.512	0.583
All (Mean)	**0.748**	**0.665**	**0.734**

B. 음성 인식 설계 및 학습 과정

1) 생활 소음 데이터셋 수집 및 전처리

16kHz 샘플링과 정규화를 수행하여 전처리를 진행하였고, 다양한 상황을 고려하여 위험 단어 리스트를 별도로 구축하였다.

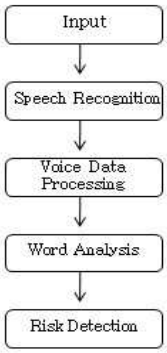

Fig. 1 음성 기반 위험 인식 모델 구조

Table 3. 위험단어 분류

분류	단어
폭력/공격	칼, 흉기, 폭행, 살해
절도/침입	도둑, 절도, 침입, 문 열어
납치/위협	납치, 감금, 협박, 위협, 도망
도움 요청	살려주세요, 도와주세요, 비명, 경찰, 신고

2) 음성 인식(Speech-to-Text)

외부 소음 및 위협 단어가 포함된 음성 파일을 WAV 형식으로 변환한 뒤, Speech-Recognition 라이브러리를 활용하여 음성을 텍스트로 변환하는 절차를 추가하였다.

3) 형태소 분석

NLTK와 Konlpy 라이브러리를 사용하여 음성 데이터에 대해 형태소 분석(Okt)을 수행하였으며, 이를 통해 '도둑', '칼', '위험'과 같은 위험 단어를 명사 기반으로 추출하였다.

4) SVM 분류기 학습

음성 특징을 추출한 후 SVM 기반 분류 모델을 활용하여 위험과 일반 소리를 학습하였다. 비명 및 사이렌(Class 1)과 일반 소리(Class 0)로 데이터를 구분하여 라벨 분포를 파악하고 스케일링을 수행하였다.

Table 4. SVM 분류 결과

항목	Precision	Recall	f1	support
Class 0	1.0	1.0	1.0	12
Class 1	1.0	1.0	1.0	18

실험결과는 200개의 오디오 샘플을 사용하여 SVM 모델을 학습하고, 학습과 검증 데이터를 7:3 비율로 분할하여 성능을 평가하였다. Precision, Recall, f1-score 값이 1.00으로 높은 예측률을 보였으나, 데이터셋 부족으로 인한 과적합을 보였고, 추가적으로 개선이 필요하다.

5) 리스크 판단 알림(Risk Detection)

실시간 입력 음성을 분석한 결과 위험 상황으로 판단될 경우, 사용자 스마트폰 또는 서버로 즉각 알림을 전송한다.

C. 멀티모달 기반 지능형 탐지 프로세스

1) 입력 데이터 수집 : 카메라와 마이크를 통해 방문자의 이미지와 음성을 실시간으로 수집하였다.

2) 이미지 기반 객체 인식 모듈 : YOLOv8m 모델을 이용하여 이미지 내 객체를 탐지하며, 칼(knife), 방망이(bat) 등 위험 객체와 우산(umbrella), 나무스틱(stick) 등 안전 객체를 구분하였다.

3) 음성 기반 위험단어 탐지 모듈 : 정상 소리와 위험 소리의 혼합을 통해 위험 상황을 실시간으로 감지한다.

4) 멀티모달 융합 로직 설계 : 객체 인식 또는 음성 분석 중 하나라도 위험 요소가 검출될 경우 이를 위협 상황으로 판단하도록 설계하였다. 또한 음성 기반 위험 단어 분석 결과와 이미지 기반 객체 인식 결과에 각각 1.6과 2.0의 가중치를 부여하여 최종 위험 점수(Risk Score)를 계산하였다.

5) 경고 및 알림 단계 : 산출된 위험 점수가 임계값(예: 0.7)을 초과할 경우, 시스템은 경고음 및 위험 알림 메시지를 즉시 출력하여 사용자가 신속히 대응할 수 있도록 하였다.

Ⅲ. 결론

본 연구는 이미지 기반 객체 인식과 음성 기반 위험 단어 탐지를 결합한 멀티모달 기반 접근 방식을 적용하여 사회적 약자의 주거 안전강화를 위한 지능형 위험 탐지 시스템을 구현하였다. 객체 인식 측면에서는 다양한 규모의 YOLOv8 모델을 활용해 위험·비위험 객체를 정확히 분류하고 실시간 탐지 가능성을 검증하였다. 경량화 모델인 YOLOv8n의 한계를 극복하기 위해 YOLOv8m 모델을 적용함으로써, 위험 객체(칼, 방망이)와 비위험 객체(우산, 나무스틱)의 분류 정밀도를 유의미하게 향상시켰다. 또한, 음성 인식 측면에서는 STT와 SVM 분류기를 융합하여 비명 및 위협 단어를 실시간으로 탐지하는 이중 분석 구조를 확립하였으며, 테스트 데이터셋에 대해 높은 정밀도를 달성하였다.

마지막으로 이미지 객체와 음성 객체의 모달리티의 결과에 가중치를 부여하는 멀티모달 융합 로직을 설계하여, 단일 센서 방식보다 신뢰성 높은 위험 탐지 프로세스를 구현함으로써 적용 가능성과 신뢰성을 확인하였다. 향후 연구에서는 객체 인식과 음성 인식 결과를 통합하는 멀티모달 기반의 위험 판단 로직의 정량적 검증이 요구된다. 아울러 본 연구가 형태적 분류에 집중했으나, 복합 객체나 순간적으로 가려진 물체에 대한 고도화된 탐지 기법 연구도 향후 확장가능한 중요한 연구 분야이다.

ACKNOWLEDGMENT

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송표준개발지원사업(RS-2024-00397768)사업으로 지원받은 연구결과임

참 고 문 헌

[1] V. Patel, S. Kanani, T. Pathak, et al., "A Demonstration of Smart Doorbell Design Using Federated Deep Learning," arXiv:2010.09687, 2020.

[2] P. Wu, J. Liu, Y. Shi, et al., "Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision," ECCV, 2020.

[3] M. Brousmiche, J. Rouat, S. Dupont, "Multimodal Attentive Fusion Network for audio-visual event recognition," Information Fusion, vol. 85, pp. 52~59, 2022.