멀티모달 정보와 언어 모델 추론을 이용한 소형 위협 객체 탐지

고현성[1], 김현수[1], 손윤식[2]
[1]동국대학교 컴퓨터·AI학과 석사과정
[2]동국대학교 컴퓨터·AI학과 교수
2025120253@dgu.ac.kr, qqaazz0222@dgu.ac.kr, sonbug@dongguk.edu

Detecting Small Threat Objects with Multimodal Evidence and LLM Reasoning

Hyunseong Ko[1], Hyunsu Kim[1], Yunsik Son[1]
[1]Dept. of Computer Science and Artificial Intelligence, Dongguk University

요 약

본 연구는 CCTV 환경에서 손에 쥐어진 소형 객체가 손에 의한 가림으로 인해 기존 객체 탐지 모델이 놓치기 쉬운 문제를 다루고, 이를 해결하기 위해 멀티모달 정보와 대규모 언어 모델(LLM)의 추론 능력을 결합한 소형 위협 객체 탐지 기법을 제안한다. 구체적으로 인체 포즈 추정을 통해 손·손목 기반 관심영역(ROI)을 설정하고, 해당 영역에서 식별된 객체 정보와 전체 프레임 정보를 함께 LLM에 입력해 추론함으로써 '객체 존재'를 판단하는 동시에, '위협 가능성'에 대한 상황적 분석을 수행한다. 통제된 데이터셋 실험 결과, 제안 기법은 부분 가림 상황에서도 '객체 존재' 판단 성능 지표가 유의하게 향상되어 국소 정보 집중과 LLM의 종합적인 추론 능력의 결합 효과를 입증하였다.

1. 서론

지능형 영상 분석 기술의 발전으로 CCTV는 범죄 예방 및 신속 대응을 위한 핵심 도구로 자리 잡았다. 특히 인공지능을 활용하여 잠재적 위협 요소를 자동으로 탐지하는 연구가 활발히 진행되고 있으며, 대부분 총기나 대형 칼과 같이 명확한 형태를 지닌 무기 객체를 인식하는 데 초점을 맞춰왔다 [1, 2]. 그러나 이러한 접근 방식은 실제 범죄 상황에서 더 빈번하게 사용될 수 있는 일상 속 소형 도구들을 간과하는 문제를 안고 있다.

실제로 위협 상황에서 사용되는 도구는 드라이버, 커터칼, 펜 등 일상에서 쉽게 접할 수 있는 소형 객체인 경우가 많다 [3]. 이러한 물체들은 손에 쥐어지는 순간 심각한 가림이 발생하고 영상 내 객체 크기가 매우 작아져 기존 탐지 모델의 성능이 급격히 저하된다. 더 큰 문제는 해당 객체가 놓인 상황과 문맥을 이해하지 못하면 위협 여부를 판단할 수 없다는 점이다. 예를 들어, 정비사가 드라이버를 사용하는 것은 정상적인 상황이지만, 괴한이 인적이 드문 곳에서 드라이버를 위협적으로 들고 있다면 이는 명백한 위협 신호이다. 이처럼 기존의 객체 탐지 기술은 시각적 형태만 인식할 뿐, 그 이면에 담긴 상황적 문맥을 해석하지 못해 오탐과 미탐을 유발하는 근본적인 한계를 지닌다.

본 연구는 이러한 한계를 극복하기 위해, 멀티모달 정보와 언어 모델 추론을 결합한 새로운 소형 위협 객체 탐지 기법을 제안한다. 이러한 접근법은 탐지 정확도를 높이는 동시에 판단의 근거까지 제시할 수 있어, 지능형 보안 시스템의 신뢰성을 한 단계 높이는 것을 목표로 한다.

2. 관련 연구

2.1. 딥러닝 기반 객체 탐지

딥러닝 기술, 특히 합성곱 신경망(CNN)의 발전은 객체 탐지 분야에 혁신을 가져왔다. R-CNN 계열의 2-stage detector와 YOLO, SSD와 같은 1-stage detector 모델들은 실시간에 가까운 속도로 높은 정확도를 달성하며 다양한 분야에 적용되었다. 특히 보안 및 안전 분야에서는 이러한 모델들을 활용하여 총기, 칼 등 특정 위협 객체를 탐지하려는 연구가 다수 진행되었다. 하지만 기존 객체 탐지 모델들은 주로 객체의 고유한 시각적 특징에만 의존하기 때문에, 형태가 비정형적이거나 손에 의해 심하게 가려진 소형 객체를 안정적으로 탐지하는 데 어

려움을 겪는다 [4, 5]. 또한, 탐지된 객체가 어떤 상황에서 어떻게 사용되고 있는지에 대한 문맥 정보를 해석하는 능력이 부재하여, 일상용품이 흉기로 사용되는 복합적인 위협 상황을 판단하는 데 명확한 한계를 보인다.

2.2. 인체 포즈 추정 기술

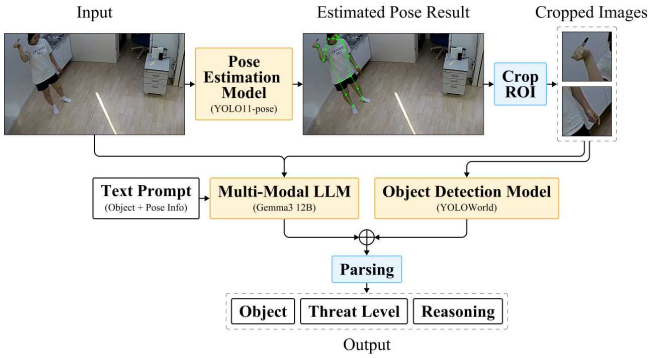
인체 포즈 추정은 영상에서 사람의 주요 관절(keypoints) 위치를 예측하여 자세나 움직임을 이해하는 기술이다. OpenPose, MediaPipe와 같은 선도적인 모델들은 다수의 사람에 대해서도 실시간으로 정확한 관절 위치를 추정하는 성능을 보여주었다. 본 연구에서는 인체 포즈 추정 기술을 행동 분석뿐만 아니라, 위협 객체를 탐지하기 위한 핵심적인 단서로 활용한다. 특히 객체를 쥔 사람의 손 위치를 특정하여 탐지 영역을 좁히고, 팔의 각도나 몸의 방향과 같은 자세 정보를 통해 '상황적 문맥'을 추출함으로써 객체 탐지만으로는 알 수 없는 고차원적인 정보를 시스템에 제공하는 역할을 한다 [6].

2.3. 시각-언어 모델

최근 자연어 처리 분야를 주도하고 있는 대규모 언어 모델(LLM)은 텍스트를 넘어 이미지, 소리 등 다른 양식(modality)의 데이터를 함께 이해하는 멀티모달(Multi-modal) 분야로 빠르게 확장되고 있다. 특히 시각-언어 모델은 이미지를 보고 관련된 질문에 답하거나(VQA), 이미지의 내용을 서술하는(Image Captioning) 등 시각 정보를 언어적으로 이해하고 표현하는 능력을 보여준다. 이러한 모델들의 핵심은 단순히 이미지를 분류하거나 객체를 탐지하는 것을 넘어, 이미지 내 객체, 인물, 배경 간의 관계를 파악하고 복합적인 상황에 대해 '추론(Reasoning)'하는 능력에 있다. 본 연구는 이러한 LLM의 추론 능력을 적극적으로 활용하여, 객체 탐지 모델과 포즈 추정 모델로부터 얻은 단편적인 시각 정보들을 종합하고, 최종적으로 주어진 상황이 실질적인 위협이 되는지에 대한 판단을 내리는 데 사용한다.

3. 제안하는 방법

본 연구에서 제안하는 소형 위협 객체 탐지 기법은 영상에서 멀티모달 정보를 단계적으로 추출하고, 이를 언어 모델의 추론 능력과 결합하여 최종적으로 상황의 위험도를 판단한다. 전체 과정은 그림 1과 같이 크게 4개의 단계로 구성된다.



(그림 1) 제안하는 기법의 전체 단계

3.1. 손 위치 특정을 위한 포즈 추정

전체 영상 프레임에서 소형 객체를 직접 탐지하는 것은 비효율적이며 정확도가 낮다. 본 연구에서는 '위협이 되는 소형 객체는 주로 사람의 손에 쥐어져 있다'는 강한 사전 지식을 활용한다. 이를 위해 먼저 사전 학습된 인체 포즈 추정 모델(YOLO11-pose)를 활용하여 영상 내 사람의 주요 관절(Keypoints)을 추출한다 [7]. 추출된 관절 정보 중에서 왼쪽 손목(L-Wrist)과 오른쪽 손목(R-Wrist)의 좌표를 식별하여, 후속 단계에서 객체 탐지를 수행할 영역을 특정하기 위한 핵심 위치 정보로 사용한다. 이 접근법은 광범위한 탐색 영역을 의미있는 국소 영역으로 좁혀줌으로써, 계산 효율성과 탐지 정확도를 동시에 향상시키는 효과를 가져온다.

3.2. 관심 영역 설정 및 객체 탐지

이전 단계에서 얻은 손목 좌표를 기준으로, 정사각 형태의 관심 영역(Region of Interest, ROI)을 설정하고 원본 이미지에서 해당 영역을 크롭한다. 이 과정은 배경 등 불필요한 정보를 제거하고 손과 그 주변의 픽셀 정보에 집중하는 효과를 주어, 저해상도의 소형 객체를 더 명확하게 분석할 수 있게 한다. 이후, 사전 학습된 객체 탐지 모델(YOLO-World)을 크롭된 ROI 이미지에 적용하여 손에 쥐어진 물체의 클래스와 정확한 위치를 식별한다 [8]. 전체 프레임이 아닌 특정 ROI 내에서만 탐지를 수행함으로, 작은 객체에 대한 탐지 민감도를 크게 높일 수 있다.

3.3. LLM 추론을 위한 정보 구성 및 프롬프트 생성

다음으로, 앞선 단계들에서 추출된 정보를 멀티모달 LLM이 이해하고 추론할 수 있는 형태로 구성한다. LLM에 장면 전체를 보여 주는 이미지, 손목 좌표를 기준으로 크롭된 이미지들, 그리고 모델이

따라야 할 텍스트 프롬프트를 함께 전달한다. 텍스트 프롬프트는 먼저 모델의 역할을 “전역 장면과 손 주변 단서를 함께 해석하는 이미지 분석 보조자”로 명시한다. 이어서 분석 규칙을 제시해 시나리오에 맞춰 정한 목표 물체 목록만 인식 대상으로 허용하고 그 밖의 물체는 전부 무시하도록 제한한다. 또한 장면 맥락과 손 영역 증거를 종합하여 '객체 존재 유무(Yes/No)'를 1차적으로 판단하도록 요구하고, 부가적으로 전반적 위험도를 0부터 10 사이의 정수로 평가하도록 요청하며, 불확실할 때는 과장 없이 보수적으로 판단하도록 톤을 안내한다. 이렇게 전역 문맥과 손 주변 단서를 한 프롬프트 안에서 결합하고, 인식 범위와 출력 형식을 미리 고정해 두면, 손에 든 소형 물체의 존재 유무 판별과 장면 기반 상황 분석을 단 한 번의 LLM 호출로 안정적으로 수행할 수 있다.

<표 1> LLM 입력 데이터

구분	입력 항목	설명
이미지	전체 프레임	사람의 전반적인 자세, 주변 환경 등 전체적인 상황의 문맥 정보를 제공
	크롭된 ROI	손에 쥐어진 객체를 명확하게 보여주어 객체 식별의 정확도를 높임
텍스트	프롬프트	추출된 정보를 바탕으로 LLM의 역할을 정의하고 원하는 결과물의 형식을 지정하는 질의를 생성

3.4. 멀티모달 LLM 기반 객체 탐지 및 상황 분석

마지막으로, 3단계에서 구성된 멀티모달 입력(전체 프레임 이미지, 크롭된 ROI 이미지, 텍스트 프롬프트)을 멀티모달 LLM인 Gemma 3 (12B)에 전달한다. 이 단계에서 LLM은 두 가지 핵심적인 역할을 동시에 수행한다[9]. 첫째, 소형 객체 존재 유무 판별(재확인) 역할이다. LLM은 객체에 집중된 크롭 이미지와 전체 문맥을 바탕으로, 손에 의한 가림 등으로 인해 초기 탐지 모델이 놓쳤을 수 있는 객체의 존재 여부를 다시 한번 정밀하게 판단한다. 이는 초기 객체 탐지 모델에서 발생할 수 있는 오류를 보정하는 효과를 가진다.

둘째, 종합적인 상황 분석 역할이다. LLM은 전체 프레임 이미지를 통해 인물의 전반적인 자세, 주변 배경, 다른 객체와의 상호작용 등 폭넓은 문맥 정보를 파악한다. 최종적으로 LLM은 이 두 가지 분석 결과를 종합하여, '객체 존재 유무'에 대한 최종 판단을 내리고, 부가적으로 주어진 상황이 실질적인 위협이 되는지에 대한 분석을 제공한다. 이 과정은 위협도 점수와 함께 "크롭 이미지에서 '드라이버'로 재확인된 객체를 든 사람이 공격적인 자세를 취하고 있음"과 같이 구체적이고 설명 가능한 근거를 함께 제공하여 시스템의 판단 과정을 투명하게 보여줄 수 있다.

4. 실험 및 결과

제안하는 기법의 성능을 객관적으로 검증하기 위해, 통제된 환경에서 자체 구축한 영상 데이터셋을 기반으로 비교 실험을 진행하였다. 실험의 핵심 목표는 제안하는 기법이 기존 단일 객체 탐지 모델 대비, 손에 쥐어져 가림이 발생한 소형 객체를 얼마나 더 정확하게 탐지하는지 입증하는 것이다.

4.1. 테스트 데이터셋 구축

드라이버나 커터칼과 같은 실제 도구를 이용한 위협 상황 데이터는 구축 과정에서 안전상의 이슈가 발생할 수 있으며, 다양한 시나리오를 재현하기에 어려움이 따른다. 이러한 현실적인 제약을 해결하기 위해, 본 연구에서는 소형 위협 객체의 특성(작은 크기, 손에 의한 가림)을 공유하면서도 안전한 실험이 가능한 '펜(Pen)'을 대표 객체로 선정하였다. 테스트 데이터셋은 펜을 들고 있는 상황을 나타내는 ‘소형 객체를 들고 있음’ 클래스와 펜을 들고 있지 않은 ‘소형 객체를 들고 있지 않음’ 클래스로 구성하였다. 각 클래스에 2,248장의 이미지를 할당하여, 총 4,496장의 이미지로 테스트 데이터셋을 구축하였다.

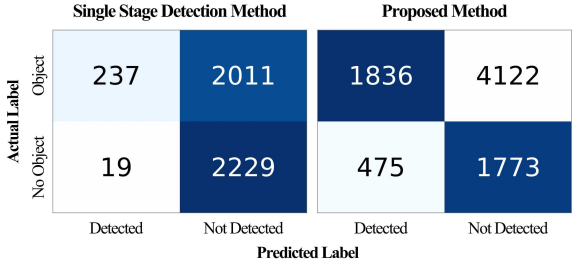
4.2. 평가 지표 및 비교 기법

모델의 성능은 소형 객체를 얼마나 정확하게 탐지하는지를 기준으로 평가하였으며, 이를 위해 탐지 정확도(Accuracy)와 F1-Score를 핵심 성능 지표로 사용하였다. 제안 기법의 유의미성을 검증하기 위해 다음과 같이 두 가지 모델을 비교하였다.

<표 2> 비교를 위한 기법별 모델 구성

구분	모델 구성
단일 단계 탐지 기법	YOLO-World
제안 기법	YOLO-World + YOLO11-pose + Gemma 3

4.3. 실험 결과 및 분석



(그림 2) 기법 별 테스트 결과

<표 3> 모델별 탐지 성능

구분	정확도(%)	F1-Score
단일 단계 탐지 기법	54.85	0.1893
제안 기법	80.27	0.8054

그림2와 표3은 두 가지 소형 객체 탐지 기법의 성능 측정 결과이다. YOLO-World 모델만을 사용한 단일 단계 탐지 기법은 손에 쥔 펜을 탐지하는 과정에서, 배경의 복잡한 요소나 대상 객체를 찾아내지 못하는 경우가 많았다. 그 결과 정확도 54.85%, F1-Score 0.1893이라는 낮은 성능을 기록했다.

반면, 본 연구에서 제안하는 기법은 불필요한 배경 정보를 효과적으로 제거하고 손에 쥔 객체에 대한 탐지 집중도를 높였다. 그 결과 정확도 80.27%, F1-Score 0.8054라는 월등히 높은 성능을 달성하여, 제안하는 기법이 손에 쥔 소형 객체 탐지 정확도를 크게 향상시킬 수 있음을 입증하였다.

5. 결론

본 연구는 기존 객체 탐지 기술이 소형 객체의 가림 및 문맥 인지에 취약한 한계를 극복하기 위해, 멀티모달 정보와 대규모 언어 모델(LLM)의 추론 능력을 결합한 탐지 기법을 제안하였다. 제안 기법은 인체 포즈 추정으로 손 주변의 주의 집중 영역을 설정하고 , 해당 영역의 국소 정보와 장면 전반의 문맥을 LLM으로 종합 분석한다. 실험 결과, 이러한 접근법은 단일 탐지기 대비 손에 쥔 소형 객체의 존재 여부를 판단하는 정확도를 현저히 향상시켰으며 , 이는 제안 기법이 심한 가림 상황에서도 탐지 성능을 크게 높일 수 있음을 입증한다.

본 연구의 기여는 이처럼 가림 상황에서도 소형 객체를 탐지하는 실증적 성능을 확보한 것과 더불어, 탐지 결과를 바탕으로 상황 맥락까지 이해하는 LLM 추론 프레임워크의 가능성을 제시한 데 있다. 이는 향후 CCTV 기반 범죄 예방 시스템이 단순 객체 탐지를 넘어 고차원적인 위협 상황을 인지하는 데 핵심적인 초석이 될 것으로 기대한다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2025-2020-0-01789). 이 성과는 정부(과학기술정보통신부)의 재원으로 과학기술사업화진흥원의 지원을 받아 수행된 연구임(과제번호 RS-2025-02412990). 이 성과는 정부(과학기술정보통신부)의 재원으로 과학기술사업화진흥원의 지원을 받아 수행된 연구임(2710086167).

참고문헌

[1] Seo, A.; Woo, S.; Son, Y., Enhanced Vision-Based Taillight Signal Recognition for Analyzing Forward Vehicle Behavior, Sensors, 24, 16, 5162, 2024.

[2] Santos, T.; Oliveira, H.; Cunha, A., Systematic review on weapon detection in surveillance footage through deep learning, Computer Science Review, 51, 100612, 2024. (article no.)

[3] Office for National Statistics, Police recorded offences involving knives or sharp instruments: methodology changes, Newport (UK), ONS, 2021.

[4] Seo, A.; Jeon, H.; Son, Y., Robust prediction method for pedestrian trajectories in occluded video scenarios, Soft Computing, 29, 4449 - 4459, 2025.

[5] Ouardirhi Z.; Mahmoudi S. A., Zbakh M., Enhancing Object Detection in Smart Video Surveillance: A Survey of Occlusion-Handling Approaches, Electronics, 13, 3, 541, 2024.

[6] Velasco-Mata A., Ruiz-Santaquiteria J., Vallez N., Deniz O., Using human pose information for handgun detection, Neural Computing and Applications, 33, (issue n/a), 17273 - 17286, 2021.

[7] Maji, D.; Nagori, S.; Mathew, M.; Poddar, D., YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss, CVPR Workshops (ECV), New Orleans, 2022, 2637 - 2646.

[8] Cheng, Tianheng; Song, Lin; Ge, Yixiao; Liu, Wenyu; Wang, Xinggang; Shan, Ying, YOLO-World: Real-Time Open-Vocabulary Object Detection, CVPR, Seattle, 2024, pp. 16901 - 16911.

[9] Gemma Team, Gemma 3 Technical Report, arXiv preprint arXiv:2503.19786, 2025.