운동정보와 의미정보 융합 CCTV 영상 내 이상 탐지에 관한 연구

홍준호, 김동성*
숭실대학교

hgoon312@naver.com, *dongsung@ssu.ac.kr

Hong Jun Ho, Kim Dong Sung*
Soongsil Univ.

요 약

기존 이상행동 탐지는 대규모 데이터 학습이 필수적이며 이를 보완하는 VLM(Vision-Language Model)인 CLIP[1]은 정지 이미지를 기반으로 실행되어 동적 행동을 파악하는데 제약이 있다. 본 연구에서는 CLIP 을 기반으로 한 의미론적 추론 정보와 객체 추적에서 생성되는 물리적 정보(속도, 크기, 픽셀 모션)와 시간적 스무딩을 결합한 뉴로-심볼릭 접근법을 제안한다. CUHK Avenue[2] 데이터셋 실험 결과 추가 학습 없이도 AUC 79.57%를 달성하였다. 이는 학습에 소요되는 시간과 비용을 '0'로 절감하면서 텍스트 수정만으로 즉시 적용할 수 있는 범용성을 보이며, 실용적 관제 솔루션의 가능성을 입증한다.

Ⅰ. 서 론

공공 안전 및 보안 등을 목적으로 CCTV 설치가 폭발적으로 증가하고 있으며 효율적이고 정확한 관제를 위하여 VAD(Video Anomaly Detection)은 필수적이다. 그러나 해당 분야에서 사용되고 있는 AutoEncoder[3]와 같은 모델을 사용한다고 했을 때, 지속적인 데이터 수집 및 학습으로 인하여 상당한 시간과 비용이 소비될 것이며 이는 CCTV 장소의 추가되거나 환경 변화 시 반복될 것이다.
최근 OpenAI의 CLIP와 같은 비전-언어 모델(Vision-Language-Model)은 자연어로 구성된 텍스트와 이미지를 매칭하여 별도의 데이터 학습 없이 이상 행동 탐지가 가능하다. 그러나 CLIP 은 단일 이미지를 분석하므로 시간의 흐름에 따라 맥락을 고려해야 하는 동작을 인지하기 어렵다.
본 논문에서는 딥러닝 모델의 분류 능력과 물리적 움직임에 대한 규칙 기반 정의를 결합한 하이브리드 프레임워크를 제안한다. YOLOv8[4]와 BoTSORT[5]를 통해 객체의 동적 정보와 CLIP 을 통한 대상의 의미 정보를 융합함으로써, 신경망 모델의 시공간적 약점을 보완하고 이상행동 탐지의 정확도를 높이고자 한다.

Ⅱ. 본론

(1) 데이터셋
본 연구를 위해서는, 다양한 형태의 객체도 중요하지만 CCTV 의 시점에서 대상을 관측한 데이터가 특히 중요하다고 할 수 있다. 본 연구에서는 CUHK Avenue 를 사용하였다. 특히 본 연구는 Zero-Shot 환경을 전제로 하므로, Training 데이터셋 없이 Testing 데이터셋 만을 사용하였다. 해당 데이터셋은 달리기, 가방 던지기, 종이 던지기, 제자리 춤추기 등 복합적인 유형의 이상행동을 다수 포함하여 단순 객체탐지가 아닌 딥러닝 모델의 구분 능력과 물리적 정보를 이용한 보정, 시공간 스무딩(Temporal Smoothing) 기법을 모두 필요로 하여 이들의 효과를 입증하는 데 최적화되어 있다.

(2) 행동 설정
(1) 텍스트 프롬프트 구성 (Text Prompt Configuration)
본 연구에서는 CLIP 모델의 제로샷 분류를 위해, 정상 행동과 비정상 행동을 자연어 문장(Text Prompt)으로 정의하였다. 특히 비정상 행동은 단순한 움직임뿐만 아니라, 특정 방향성이나 상호작용이 포함된 복합 행동을 포괄할 수 있도록 세분화하여 구성하였다.

정상 행동 (Normal Group): 보행자의 일상적인 패턴

"a person walking naturally along the street"
"a person walking away from the camera" (카메라 등지고 걷기)
"a person carrying a bag or backpack"
"people commuting typically"

비정상 행동 (Abnormal Group)

동적 위협 행동 (Dynamic Actions):

"a person running fast" (달리기)
"a person throwing a bag or object" (물건 투척)
"a person throwing papers" (전단지/종이 투척)
"a person dancing or performing" (춤)
"a person fighting or hitting" (폭력)

비정상적 이동 및 접근 (Spatial/Directional Anomalies):

"a person walking towards the camera" (카메라로 접근)
"a person moving strangely" (이상한 움직임)

(3) 모델 파이프라인
본 연구의 파이프라인의 흐름은 다음과 같다.

Step 1. 객체 탐지, 추적 및 ROI 최적화(Detection, Tracking, ROI Refinement)

영상 내에서 YOLOv8 을 이용하여 사람을 탐지하고, 나아가 각 사람별로 BoTSORT 를 사용하여 각 사람의 ID 를 부여하며 각 ID 별 상태 정보(좌표, 이동 속도, 대상의 가로/세로 비율 등)를 지속적으로 획득한다.

이 때, 식별된 사람 기준 타이트한 바운딩 박스가 생성되는데 이 경우 물건 투척(a person throwing a bag or object)와 같은 행동이 발생했을 때, 행동의 문맥이 손실되는 문제가 발생한다. 이에 따라 본 연구에서는 적응형 상단 크롭 이미지를 새로 생성하게 되며 이 이미지의 영상 내 좌표는 다음과 같이 설정된다.

$[x_1, y_1, x_2, y_2]$: 기존 사람의 바운딩 박스 좌표

$$y_1^{'} = max(0, y_1 - 0.2h)$$

$$h = y_2 - y_1$$

$$x_1' = x_1, \quad x_2' = x_2, \quad y_2' = y_2$$

Step2. 이원화된 하이브리드 분석

2.1 의미론적 분석
Tracker 로부터 추출된 각 사람의 ROI 이미지를 딥러닝 모델(CLIP)에 적용하여 대상의 시각-언어적 상태를 분석하여 의미론적 분석을 실시한다.
여기서는 CLIP 모델을 사용하여 ROI 이미지 벡터와 대상의 이미지가 달리기, 싸우기 등 주어진 텍스트와 코사인 유서도를 계산하여 얼마나 유사한지 여부를 계산하게 된다.

2.2 물리적 분석
Tracker에서 생성된 대상의 상태 정보(대상 사람의 좌표, 이동 속도, 박스 크기, 이전 프레임 정보)를 분석, 대상 사람의 사이즈가 지나치게 커지는 경우(화면 점유율 40% 이상), 수직 이동 속도가 높은 경우, 각 바운딩 박스 픽셀 내부의 픽셀 차분이 급격히 커질 경우 이상 행동으로 간주하게 된다.

Step3. 통합 및 점수 설정
우선 속도를 기반으로 점수를 부여하며, 속도가 느린 경우 페널티를 부여한다. 다만, 대상의 속도가 느리더라도 발작, 춤 등 이상 행동일 수 있는데 다음과 같은 로직으로 판별하여 그 문제점을 해결한다.

1. 객체의 이동 속도가 임계값 이하인가? $(v < \tau_{vel})$

2. 이전 프레임과의 픽셀 차분 평균이 기준값 이상인가? $(\Delta I > \tau_{pixel})$

1, 2 번 조건이 True 이면 제자리에서의 이상 행동으로 판단하고 속도 페널티를 적용하지 않으며 가산점을 부여하여 이상행동 가능성을 올린다. 이 때, 단순 비정상 텍스트와의 유사도($P_{abnormal}$)만으로 점수를 산출한다면, 다수의 인물 또는 배경 사물이 등장할 경우 모든 행동의 확률이 전반적으로 높게 나오는 상황이 발생한다. 따라서 점수 산출 방식은 다음과 같이 설정한다.

$$S = \max(0, P_{abnormal} - P_{normal})$$

이를 통하여 배경 또는 군집 등의 배경이 있어서 모든 점수가 높게 나오는 상황에서 무조건 비정상 확률이 높아서 나오는 오탐을 방지하도록 한다.

Step4. 점수 최종 산출
CLIP 이 프레임별로 독립적으로 점수를 산출한다. 인접한 프레임 사이에서도 실제 이상행동 발생 또는 오탐(False Positive)에 의하여 점수가 급격히 변동(Jittering)하는 현상이 발생하게 된다. 본 연구에서는 이를 지수 이동 평균(Exponential Moving Average)으로 점수의 등락을 조절하여 노이즈(오탐)이 이상 탐지에 영향을 미치는 것을 억제하고 지속적인 이상행동 패턴에 대한 탐지 능력을 강화하여 탐지의 신뢰성을 강화하였다.

Ⅲ. 결론

본 연구에서는 새로운 데이터 수집과 학습을 하지 않았으며 대신 텍스트 프롬프트와 물리적 정보를 바탕으로 한 규칙을 결합하여 하이브리드 제로샷 탐지 프레임워크를 제안하였다. 이 필수였던 기존 7 9.6%의 AUC 를 달성하였다. 성능을 SOTA[6]와 비교할 만큼 향상시키기 위하여 CLIP-Adapter[7]와 같은 어댑터 모듈을 추가하여 극소량의 데이터(10~20 장) 학습, CoOp(Context Optimization)과 같은 기법을 도입하여 연구를 보다 고도화 하고자 한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 메타버스 융합대학원의 연구 결과로 수행되었음

(IITP-2026-RS-2024-00430997).

참 고 문 헌

[1] A. RadFord et al., “Learning Transferable Visual Models From Natural Language Supervision,” in International Conference on Machine Learning (ICML), 2021, pp. 8748-8763

[2] Lu, C., Shi, J., and Jia, J ” Abnormal event detection at 150 FPS in MATLAB.” In IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013, pages 2720-2727.

[3] Viorica P˘ atr˘ aucean, Ankur Handa & Roberto Cipolla “SPATIO-TEMPORAL VIDEO AUTOENCODER WITH DIFFERENTIABLE MEMORY” 2016

[4] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[5] Nir Aharon,Roy Orfaig Ben-Zion Bobrovsky “BoT-SORT: Robust Associations Multi-Pedestrian Tracking,” 2022.

[6] H. Wang et al., "Multimodal Memory Learning for Video Anomaly Detection," in *CVPR*, 2023.

[7] P. Gao et al., "CLIP-Adapter: Better Vision-Language Models with Feature Adapters," in *International Journal of Computer Vision (IJCV)*, 2024.