CLIP 임베딩과 GRU 를 이용한 CCTV 이상 행동 감지 시스템

김현기, 윤수연*
국민대학교, *국민대학교

avalonia@kookmin.ac.kr *1104py@kookmin.ac.kr

CLIP-GRU: A Hybrid Framework for Improving Anomaly Detection in Surveillance Environments

Hyun Ki Kim, *Soo-Yeon Yoon
Kookmin Univ, *Kookmin Univ.

요 약

최근 지능형 영상 관제 시장에서는 별도의 추가 학습 없이 다양한 도메인에 적용 가능한 CLIP 기반의 zero shot anomaly detection 시스템이 주목받고 있다. 이러한 시스템은 Detector-Tracker-CLIP 파이프라인을 통해 객체의 정적인 특징을 빠르게 파악할 수 있다는 편의성을 제공하지만, 폭행, 실신과 같이 동적인 움직임과 맥락 이해가 필수적인 액션 중심의 이상 행동에서는 낮은 정확도를 보인다. 본 논문에서는 이러한 한계를 극복하기 위해 CLIP 의 시각적 임베딩 시퀀스를 GRU(Gated Recurrent Unit) 네트워크와 결합한 시계열 행동 판정 프레임워크를 제안한다. 실험 결과, 정적 이미지 기반의 zero shot 방식 대비 시계열 정보를 반영한 제안 모델이 복잡한 동작이 포함된 폭행, 실신 행동 탐지에서 zero shot anomaly detection 시스템보다 향상된 성능을 보임을 확인하였다.

I. 서론

최근 CCTV 통합 관제 센터는 영상 데이터를 효율적으로 처리하기 위해 자동화된 이상 행동 감지 시스템을 적용하고 있다. 데이터 라벨링 비용을 절감하고 범용성을 높이기 위해, 대규모 언어-이미지 사전 학습 모델인 CLIP 을 활용한 zero shot 기반 탐지 기법이 현업에서 널리 채택되는 추세이다. 이 방법은 Detector 와 Tracker 를 결합하여 추출된 객체 이미지를 CLIP 을 사용하여 가장 유사한 프롬프트와 매칭함으로써, 별도의 행동 학습 없이도 이상 징후를 감지할 수 있다는 강력한 편의성을 제공한다.

그러나 Detector-Tracker-CLIP 구조는 정적 이미지만 적용 가능한 구조적 한계를 지닌다. 실신과 같은 지속 상태 위주의 감지는 가능하나, 쓰러진 사람 주위 사람이 있는 경우와, 폭력처럼 역동적인 프레임 변화가 행동을 결정하는 경우들은, 단일 프레임 만으로는 해당 동작이 무엇인지 파악하기 어렵다. 예를 들어, 두 사람이 밀착해 있는 정적 이미지는 단순 접촉인지 폭력 인지 구분할 수 없으며, 이는 관제 시스템의 오탐지율을 높이는 주요 원인이 된다.

이 구조적 한계를 개선하기 위해, CLIP 임베딩에 GRU 의 시계열 모델링을 결합한 파이프라인 구조를 제안한다. 이를 통해 연산 효율성을 유지하면서도 폭행, 실신과 같은 이상 행동에 대한 탐지 성능을 향상시킴으로써, 지능형 영상 관제 시스템의 실질적인 고도화 방안의 하나를 제시하고자 한다.

II. 관련 연구

2.1. CLIP4Clip

CLIP4Clip 은 정적 CLIP 모델을 비디오-텍스트 검색 및 분류 영역으로 확장하여 시간적 유사성 보존의 중요성을 입증한 연구이다. 본 연구는 이 방식에 착안하여, 탐지된 객체의 BBox 에서 추출된 CLIP 임베딩 시퀀스가 개별 행동의 동적인 특성을 효과적으로 내포하고 있음을 전제로 시계열 분석을 수행한다.

2.2. VideoCLIP

VideoCLIP 은 대조 학습을 통해 비디오의 시간적 맥락을 고차원 벡터로 표현하는 기법을 제안하였다. 본 연구는 이러한 맥락 정보의 핵심 가치를 수용하되, 실시간 관제 환경에 최적화된 시스템 구축을 위해 기존의 무거운 Transformer 구조 대신 경량화된 GRU 모델을 결합하여 연산 효율성과 탐지 성능을 동시에 확보하였다.

2.3. Real-world Anomaly Detection in Surveillance Videos

CCTV 환경의 이상 행동 감지는 실제 현장의 노이즈와 데이터 불균형 문제를 해결하는 것이 핵심이다. 본 연구는 선행 연구에서 정의된 이상 행동의 물리적 특성을 바탕으로 클래스를 정의하였으며, 단순 이상치 탐지를 넘어 폭력과 실신을 구체적으로 식별함으로써 지능형 관제 시스템의 실무적 효용성을 강화하였다.

III. 실험

3.1 실험 데이터 및 파이프라인

실험에 사용한 데이터는 AI Hub 의 이상행동 감지 CCTV 영상 데이터를 기반으로 모델을 학습 및 평가하였다. 해당 데이터셋은 실제 환경의 노이즈와 다양한 조도, 각도, 해상도 조건에서 촬영된 이상행동 및 정상행동 영상을 포함하고 있어 모델의 강건성 검증에 적합하다. 이 데이터셋에서 사회적 중요도가 높은 폭행, 실신과 정상 동작의 3 개의 분류 클래스를 지정했으며, 각 클래스마다 영상 개수를 절반으로 나누어, 한 그룹은 학습에 사용했고, 남은 그룹은 추론에 사용했다.

실험에 사용한 모델의 종류들은 Detector 는 Yolov8m, Tracker 는 ByteTrack, 특징 추출은 CLIP(ViT-B/32)

모델을 활용하였으며, 이 시퀀스 데이터는 실시간 관제 환경을 고려하여 파라미터 수가 적고 연산 효율이 높은 경량

네트워크인 GRU 를 거쳐 최종적으로 행동 카테고리를 분류한다. 또한, 기존 시스템인 파이프라인에 실험 모델과 동일한 모델들을 적용하여, 성능을 대조한다.

3.2 실험 결과

객체별 정답 결과는 없지만, 영상에서 특정 구간동안 이상 행동을 수행한 기록이 있는 데이터셋 annotation 구조를 고려하여, frame 이미지 전체를 zero shot detection 했을 때, frame 마다 잡힌 객체들 중 가장 높은 이상행동 결과 수치를 지닌 클래스를 대표로 지정했을 때를 비교군으로 삼고 AUC 정량평가를 수행하였다.

[표 1]. AUC 정량평가

파이프라인	AUC
zero shot detection (frame)	51.05%
zero shot detection (bbox)	62.95%
CLIP GRU(bbox)	**69.52%**

[그림 1]은 정성적 분석 결과로서, 왼쪽은 zero shot 방식이며, 객체가 쓰러져 있는 상태에 집중하여 '실신'으로 추론했으나, 제안하는 CLIP-GRU 구조는 객체를 '폭행'으로 추론하였다. 기존의 zero shot 방식은 정적 이미지만 분석하여, 옆사람에게 의지하는 상태로 추론하지만, 제안 모델은 이전 프레임의 동작을 참고했기 때문이다.

[그림 1]. 시각화 결과: zero shot (좌), CLIP-GRU (우)

3.3 성능 평가 비교 분석

실험 결과, 제안하는 CLIP-GRU 구조는 기존 zero shot 방식보다 약 6.57%p 향상된 69.52% AUC 를 달성했다. 이는 정적 이미지의 시각 특징에만 의존하던 기존 시스템의 한계를 시계열 맥락의 결합을 통해 극복할 수 있음을 보인다. 특히 [그림 1]을 참고하면, zero shot 방식은 바닥에 쓰러져 있는 이미지만 보고 이를 실신으로 오판하는 경향이 짙으나, 제안 모델은 8 프레임의 시퀀스 임베딩을 통해 폭력이라는 동적 맥락을 확보함으로써 오판을 효과적으로 정정했다.

다만, CLIP-GRU 구조는 Detector-Tracker-CLIP-GRU 로 이어지는 직렬형 파이프라인 구조를 채택하고 있어, 상위 단계인 Detector 나 Tracker 의 성능에 의존하는 구조적인 취약성을 지닌다. 객체가 조도 변화나 폐쇄 등의 문제로 detector 에 탐지되지 않거나, Tracking ID 가 부여되지 않는 경우, 해당 프레임의 데이터가 CLIP 특징 추출 단계로 전달되지 못해 GRU 가 시계열분석을 할 수 없는 문제가 있다. 실험 과정에서 관찰된 성능 저하는 이러한 탐지 및 추적 누락이 이후 모델에 영향을 주는 cascading error 로 분석되며, AUC 수치가 저하되는 요인으로 작용했다.

IV. 결론 및 향후 연구

본 논문에서는 CLIP 의 범용적 이미지 임베딩과 GRU 의 시계열 모델링을 결합하여, 기존 zero shot 기반 이상 행동 감지 시스템이 가진 정적 분석의 한계를 해결하고자 하였다. 실험을 통해 69.52%의 AUC 를 달성함으로써 시계열 정보가 폭행 및 실신과 같은 역동적인 동작 판별에 핵심적인 역할을 수행함을 정량적으로 입증하였다. 비록 파이프라인 구조상 상위 모델의 탐지 성능에 종속되는 한계가 존재하나, 본 연구가 제시한 하이브리드 프레임워크는 경량화와 성능 사이의 적절한 균형점을 제시했다는 점에서 의의가 있다.

이 실험은 실시간 환경을 가정한 경량화 아키텍처의 PoC 단계로, 향후 실시간성을 갖추면서도 실무에 가용 할 수 있는 수준인 AUC 80% 이상의 AUC 를 확보하기 위해 다음과 같은 고도화를 추진할 계획이다.

첫째, 현재 GRU 구조를 개선하여, 장기 시퀀스 정보를 수용할 수 있는 어텐션 메커니즘을 도입함으로써 폭행, 실신 외의 다른 이상 행동에도 적용시킬 예정이다.

둘째, 앞서 언급한 탐지 및 추적 누락 문제를 완화하기 위해 저해상도 및 야간 환경에 특화된 도메인 적응 기법과 Tracker 의 강건성을 높이는 연구를 병행할 예정이다.

후속 연구들로 부터 시스템의 신뢰도를 보완하면 지능형 영상 관제 시스템에서의 실효성을 극대화할 수 있을 것으로 보인다.

참 고 문 헌

[1] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8,” 2023. [Online]. Available: https://github.com/ultralytics/ultralytics [Accessed: May 2025].

[2] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Luo, Y. Liu, S. Zhang, Z. Liu, and J. Wang, “ByteTrack: Multi-Object Tracking by Associating Every Detection Box,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–21.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[4] Z. Luo et al., "CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval and Captioning," *arXiv preprint arXiv:2104.08860*, 2021.

[5] H. Xu, G. Ghosh, P. Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, “VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 6787–6800.

[6] W. Sultani, C. Chen, and M. Shah, “Real-world Anomaly Detection in Surveillance Videos,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6479–6488.