이상행동 인식을 위한 반자동 엔드-투-엔드 딥러닝 파이프라인 구축

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

A Semi-automated End-to-End Pipeline for Deep Learning-based Abnormal Behavior Recognition

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

요 약

지능형 영상 감시의 중요성이 커지고 있지만, 이상행동 인식 모델의 개발은 데이터 준비부터 실제 배포까지의 과정이 복잡하고 비효율적인 문제가 있다. 본 논문에서는 이러한 문제를 해결하기 위해, 영상 데이터 처리부터 모델 배포 및 테스트까지 전 과정을 체계적으로 통합하는 반자동 엔드-투-엔드 파이프라인을 제안한다. 제안하는 파이프라인은 YOLOv11과 SAM2를 이용한 자동 마스킹, Streamlit 기반의 웹 UI를 통한 반자동 라벨링, RTMO 모델 기반의 스켈레톤 추출, ProtoGCN 모델 학습, 그리고 ONNX 변환 및 Triton 서버 배포의 단계로 구성된다. 본 연구의 핵심 기여는 딥러닝 기반 자동화 모듈과 실용적인 데이터 구축 단계를 결합하여, 모델 개발의 전체 주기를 관리하는 체계적이고 재현 가능한 워크플로우를 정립함으로써 연구와 실제 운영 간의 간극을 줄였다는 데 있다.

1. 서론

현대 사회에서 CCTV의 역할이 중요해졌지만, 방대한 영상 데이터를 인간이 직접 감시하는 것은 비효율적이며 사실상 불가능하다. 이에 대한 해결책으로 폭력, 쓰러짐 등 이상행동을 자동으로 탐지하는 AI 기반 지능형 감시 기술의 필요성이 급증하고 있다. 그러나 기존 연구들은 대부분 모델의 성능 향상에만 집중할 뿐, 데이터 수집부터 실제 배포까지 이어지는 단절되고 비효율적인 개발 과정의 문제를 간과하고 있다. 이러한 개발 과정의 장벽은 연구 결과를 현장에 적용하는 데 큰 어려움으로 작용한다.

본 논문은 이러한 문제점을 해결하기 위해, 데이터 준비부터 배포까지의 전 과정을 통합하는 반자동 엔드-투-엔드(End-to-End) 파이프라인을 제안한다. 제안하는 파이프라인은 마스킹, 모델 학습 등 핵심 과정은 자동화하되, 정교한 판단이 필요한 라벨링 단계에서는 인간의 개입을 효율적으로 결합하여 데이터의 품질과 구축 속도의 균형을 맞춘다. 따라서 본 연구의 핵심 기여는 이상행동 인식 모델의 전체 개발 주기를 관리하는 체계적이고 재현 가능한 워크플로우를 정립하여, 연구 개발과 실제 운영 환경 간의 간극을 줄였다는 데 있다.

2. 관련 연구

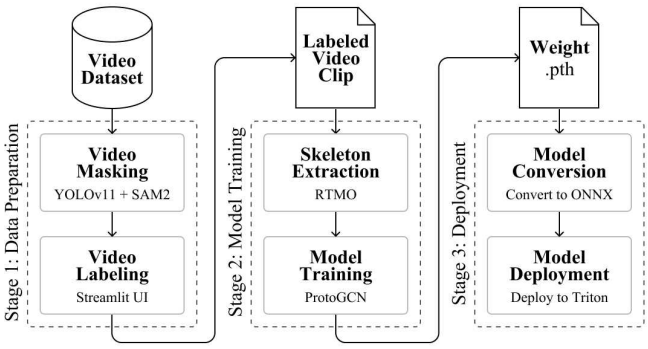
초기 행동 인식 연구는 3D CNN과 같은 RGB 영상 기반으로 발전했으나, 복잡한 배경과 높은 연산량의 한계가 있었다 [1]. 이를 극복하기 위해 2D 인체 관절 정보(스켈레톤)나 단일 영상에서 3D 인체 메쉬를 복원하는 등 인간의 핵심 자세 정보에 집중하는 연구가 활발하며, 본 연구는 이 중 스켈레톤 기반 접근법을 채택한다 [2]. 특히 ST-GCN과 같은 GCN(Graph Convolution Network)는 인체 골격을 그래프로 모델링하여 높은 성능을 보여주는 핵심 방법론이다.

모델 개발만큼 효율적인 배포도 중요하며, MLOps 분야에서는 ONNX와 Triton Inference Server를 활용한 자동화가 표준 전략이다 [3]. 하지만 기존 행동 인식 연구는 모델 제시에만 집중하고 배포까지의 전체 파이프라인을 다루는 경우는 드물다. 본 연구는 최신 모델링 기법과 검증된 MLOps 전략을 결합하여 이러한 공백을 메우고자 한다.

3. 제안하는 파이프라인

본 연구에서 제안하는 파이프라인은 원본 영상 데이터 입력부터 이상행동 예측 결과 출력까지 전

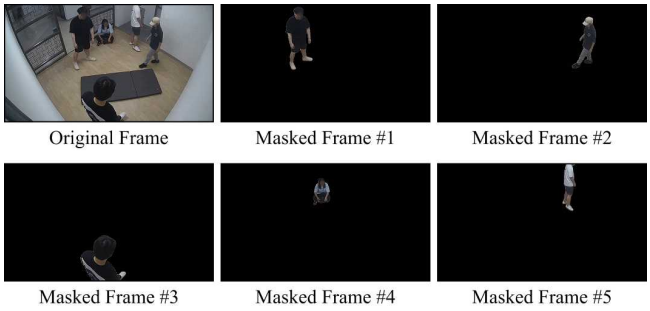
과정을 효율적으로 처리하는 것을 목표로 한다. 파이프라인은 크게 데이터 준비, 모델 학습, 그리고 모델 배포 및 테스트의 3단계로 구성되며, 각 단계는 유기적으로 연결되어 있다. 데이터 준비 단계에서는 자동화된 딥러닝 모델을 활용하여 처리 부담을 줄이되, 라벨링 과정에서 사용자의 검수를 포함하는 반자동 방식을 채택하여 데이터 품질을 확보한다. 이후 단계들은 모두 자동화되어, 학습된 모델이 최종적으로 실제 운영 환경과 유사한 서빙 환경에 배포되어 성능을 검증받게 된다. 제안하는 파이프라인의 전체 흐름은 아래 그림과 같다.



(그림 1) 제안하는 파이프라인의 전체 흐름도

3.1. 영상 마스킹

파이프라인의 첫 단계는 원본 영상에서 행동 분석에 불필요한 배경을 제거하고 특정 인물에 집중하는 것이다. 복잡한 배경이나 다른 사람의 움직임은 모델 학습에 노이즈로 작용할 수 있기 때문에, 이를 제거하기 위해 2단계 접근법을 사용한다. 먼저, YOLOv11 모델을 사용하여 각 프레임에서 모든 인물의 위치를 바운딩 박스(Bounding Box)로 식별한다 [4]. 이후, 검출된 바운딩 박스를 프롬프트로 활용하여, 정교한 인스턴스 분할이 가능한 SAM2(Segment Anything Model 2) 모델에 입력한다 [5]. SAM2는 해당 영역 내의 인물 객체에 대한 정밀한 픽셀 단위 마스크(pixel-wise mask)를 생성한다. 최종적으로, 생성된 마스크 영역을 제외한 모든 배경 픽셀을 검은색으로 처리하여 각 인물에만 초점을 맞춘 영상을 생성한다. 이 과정을 통해 모델이 순수하게 인물의 행동 패턴에만 집중하여 학습할 수 있는 환경을 제공한다.

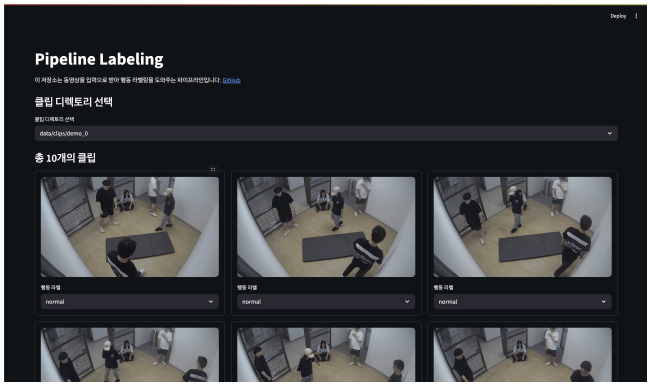


(그림 2) 영상 마스킹 결과

3.2. 영상 라벨링

정제된 영상 데이터는 모델 학습을 위해 클립 단위로 분할되고, 각 클립에 해당하는 행동 라벨이 부여된다. 이 단계는 고품질의 학습 데이터를 구축하는 데 가장 중요하지만 시간이 많이 소요되므로, 효율적인 반자동 방식을 적용한다. 이를 위해, 본 연구에서는 Python의 Streamlit 라이브러리를 활용하여 직관적인 웹 기반 라벨링 UI를 자체적으로 구현하였다.

이 웹 UI 환경에서 작업자는 지정된 폴더의 영상 목록을 확인할 수 있다. 영상을 선택하면, 시스템은 해당 영상을 일정한 시간 단위의 클립으로 자동 분하여 그리드형태로 화면에 표시한다. 작업자는 표시된 클립을 보면서 드롭다운 메뉴를 통해 간단하게 라벨을 설정할 수 있다. 이 방식은 작업자가 영상 분할이나 파일 관리와 같은 부수적인 작업 없이 라벨링 본연의 임무에만 집중할 수 있게 하여, 전체 데이터 구축 과정의 생산성과 편의성을 크게 향상시킨다.



(그림 3) 구현된 라벨링 웹 UI

3.3. 스켈레톤 추출

라벨링이 완료된 영상 클립은 GCN 계열의 행동 인식 모델이 직접 처리할 수 없는 픽셀 데이터이므로, 모델의 입력 형식에 맞는 그래프 구조의 특징

데이터로 변환하는 과정이 필수적이다. 이를 위해, 본 파이프라인에서는 실시간 고성능 포즈 추정 모델인 RTMO(Real-Time Multi-Person Pose Estimation Transformer)를 사용하여 각 프레임 속 인물의 주요 신체 관절의 2D 좌표를 추출한다 [6].

이렇게 추출된 프레임별 관절 좌표의 시퀀스가 바로 '스켈레톤 데이터'이며, 이는 행동 인식을 위한 핵심 특징으로 사용된다. 이 데이터는 인물의 형태와 움직임에 대한 핵심 정보만을 압축하고 있어, 배경이나 조명 같은 불필요한 정보 없이 모델이 행동 자체에 집중하여 학습할 수 있도록 돕는다.

3.4. 행동 인식 모델 학습

이 단계에서는 스케레톤 데이터와 해당 데이터의 라벨을 하나의 쌍으로 묶은 데이터셋을 이용하여 행동 인식 모델을 학습시킨다. 본 파이프라인은 스켈레톤 데이터를 입력으로 받는 다양한 GCN 계열의 행동 인식 모델과 호환되도록 설계되어, ST-GCN, 2s-AGCN 등 여러 SOTA(State-of-the-art) 모델과 유연하게 적용할 수 있는 확장성을 가진다.

본 연구에서는 다양한 모델 중에서도 프로토타입 기반 학습(Prototype-based learning)을 통해 적은 데이터로도 높은 일반화 성능을 보이는 것으로 알려진 ProtoGCN 모델을 채택하여 실험을 진행하였다 [7]. 학습은 교차 엔트로피 손실 함수(Cross-Entropy Loss)를 최소화하는 방향으로 진행되며, 학습이 완료되면 특정 스켈레톤 시퀀스가 입력되었을 때 어떤 행동에 해당하는지 분류할 수 있게 된다.

3.5. 모델 변환

학습이 완료된 딥러닝 모델(PyTorch의 pth파일)은 특정 프레임워크에 종속되어 있어 다른 환경에서 사용하기 어렵고, 학습에 사용된 불필요한 정보들로 인해 파일 크기가 크다. 따라서 실제 운영 환경에 효율적으로 배포하기 위해 모델을 표준화되고 최적화된 형식으로 변환해야한다. 본 파이프라인에서는 ONNX 형식을 사용한다. ONNX는 다양한 프레임워크와 하드웨어에서 호환되는 중간 표현(Intermediate Representation)으로, 모델을 ONNX로 변환하면 프레임워크 종속성에서 벗어날 수 있으며, 추론 속도 향상과 같은 다양한 최적화를 적용할 수 있는 장점이 있다.

3.6. 모델 배포 및 테스트

최종적으로, 변환된 ONNX 모델을 실제 서비스와 유사한 환경에 배포하여 성능을 검증한다. 이를 위해, 다중 모델 서빙 및 높은 처리량에 특화된 NVIDIA Triton Inference Server를 테스트 환경으로 구축한다. ONNX 모델을 Triton 서버에 로드하면, 서비는 외부에서 추론 요청을 받을 수 있는 표준 API 엔드포인트(Endpoint)를 자동으로 생성한다. 이후, 테스트 클라이언트에서 새로운 스켈레톤 데이터를 API를 통해 서버로 전송하고, 서버가 행동 예측 결과를 반환하는 전체 과정을 테스트한다.

4. 실험 및 결과

4.1. 평가 지표

제안하는 파이프라인의 성능을 종합적으로 평가하기 위해, 본 연구에서는 두 가지 측면의 지표를 사용한다. 첫째는 학습된 행동 인식 모델 자체의 분류 성능이며, 둘째는 Triton 서버에 배포된 모델의 추론 서빙 능력이다.

테스트 데이터셋에 대한 모델의 행동 분류 정확도를 평가하기 위해 다음과 같은 표준 분류 지표를 사용한다. 특히, 데이터셋 내 클래스 불균형 문제를 고려하여 정확도(Accuracy)뿐만 아니라 정밀도(Precision), 재현율(Recall), F1-Score를 함께 측정하여 모델 성능을 다각적으로 분석한다.

Triton Inference Server에 배포된 ONNX 모델이 실제 운영 환경에서 얼마나 효율적으로 작동하는지 평가하기 위해 응답 지연 시간과 처리량을 측정한다.

4.2. 데이터셋

모델 학습과 평가를 위해 자체 구축한 15초 분량 영상 1,821개로 구성된 자살·자해 행동 데이터셋과 10초 분량 영상 100개로 구성된 평가용 데이터셋을 사용하였다.

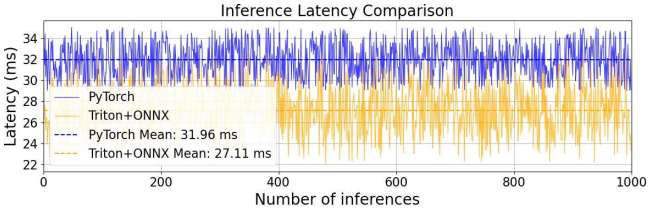
<표 1> 데이터셋 구성

Dataset / Class	Selfharm Dataset		Test Dataset
	Train	Validation	Test
Normal	803	88	50
Selfharm	835	95	50
Total	1,638	183	100

4.3. 실험 결과 및 분석

<표 2> 모델 성능 평가

Accuracy	Precision	Recall	F1-Score
0.8972	0.9140	0.8842	0.8988



(그림 4) 추론 환경 별 추론 지연 시간

<표 3> 추론 환경 별 처리량

Inference Environment	Throughput (IPS)
PyTorch	31.289
Triton+ONNX	36.886
Triton+ONNX Dynamic Batching	247.136

5. 결론

본 논문에서는 이상행동 인식을 위한 모델 개발의 전 과정을 효율적으로 통합 관리하는 반자동 엔드-투-엔드 파이프라인을 제안하고 구현하였다. 기존 연구들이 대부분 모델 성능 자체에만 집중하여 실제 개발 과정에서 발생하는 데이터 처리 및 배포의 어려움을 간과하는 문제를 해결하고자 하였다. 제안하는 파이프라인은 YOLOv11과 SAM2를 이용한 영상 마스킹, Streamlit 기반의 반자동 라벨링, RTMO를 통한 스켈레톤 추출, ProtoGCN 모델 학습, 그리고 ONNX 변환 및 Triton 서버 배포에 이르는 전 과정을 체계적으로 연결한다. 본 연구의 가장 큰 의의는 단순히 특정 모델의 성능을 개선한 것이 아니라, 딥러닝 기반의 자동화 모듈과 실용적인 인간 상호작용을 결합하여 연구 개발부터 실제 운영까지의 간극을 줄이는 체계적이고 재현 가능한 워크플로우를 정립했다는 점이다.

향후 본 파이프라인은 완전 자동화 및 확장성 증대를 목표로 더욱 발전시켜 나갈 수 있다. 이를 위해, 현재 반자동으로 이루어지는 라벨링 과정에 능동 학습(Active Learning) 기법을 도입하여 효율을 극대화하고, 실제 영상에서 빈번하게 발생하는 가림(occlusion) 상황에서도 강건한 예측이 가능하도록 모델을 개선하는 연구가 필요하다 [8]. 더 나아가, 다양한 최신 모델의 실험 및 관리를 지원하도록 파이프라인의 확장성을 높이는 것 또한 중요한 향후 과제이다.

ACKNOWLEDGMENT

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). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2025-2020-0-01789)

참고문헌

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016, pp. 770-778.

[2] H. Kim and Y. Son, "Generating Multi-View Action Data from a Monocular Camera Video by Fusing Human Mesh Recovery and 3D Scene Reconstruction," Applied Sciences, vol. 15, no. 19, p. 10372, 2025.

[3] Bai, J., et al., "ONNX: Open Neural Network Exchange," GitHub repository, GitHub, 2017.

[4] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," arXiv preprint arXiv:2410.17725, 2024.

[5] N. Ravi, V. Gabeur, Y. Hu, et al., "SAM 2: Segment Anything in Images and Videos," arXiv preprint arXiv:2408.00714, 2024.

[6] S. Luo, C. Wang, Y. Geng, H. Zhang, and W. Yang, "RTMO: Real-Time Multi-Person Pose Estimation based on Transformer," arXiv preprint arXiv:2310.05972, 2023.

[7] Z. Li, X. He, Y. Zhang, and X. You, "Prototype-based Graph Convolutional Network for Skeleton-based Action Recognition," in Proceedings of the 29th ACM International Conference on Multimedia (MM '21), Chengdu, China, 2021, pp. 317-325.

[8] A. Seo, H. Jeon, and Y. Son, "Robust prediction method for pedestrian trajectories in occluded video scenarios," Soft Computing, vol. 29, pp. 4449 - 4459, 2025.