자세 기반 위협 동작 인식을 위한 경량 모델 설계

류문욱[1,†], 권준형[2,†], 이장원[2,*]

[1]한국전자통신연구원, [2]성균관대학교

moonwook@etri.re.kr, {ludin9, leejang}@skku.edu

Lightweight Anomaly Detection via Human Pose Estimation

Moonwook Ryu[1,†], Joonhyung Kwon[2,†], Jangwon Lee[2,*]

[1]Electronics and Telecommunications Research Institute, Republic of Korea, [2]Sungkyunkwan Univ.

요 약

최근 영상 관제 시스템의 증가로, 관제 요원의 모니터링 대수가 급속도로 증가하고 있으며 이를 위해, 지능형 영상 관제 시스템이 빠르게 발전하고 있다. 하지만 높은 성능을 가진 동작 인식 기술은 무거우며, 가려짐이 심한 환경에서는 인식이 어렵기에 복잡한 모델을 현장에 바로 적용하기에는 어려운 상황이다. 이를 해결하기 위해 본 논문은 위협 동작을 인식하기 위한 실시간으로 구동이 가능한 경량 모델을 제안하며, 효율적으로 위협 동작을 탐지할 수 있는 가능성을 제시하였다.

Ⅰ. 서 론

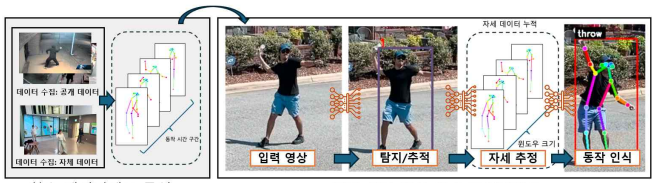
최근 범죄 및 재난·안전사고 예방을 위해 영상 관제 시스템이 빠르게 증가하고 있으나, 관제 인력은 부족한 상황이다. 서울연구원 보고서에 따르면 공공기관에 설치된 CCTV는 2018년 기준 1,032,879대가 설치 운영되고 있으며, 서울시 자치구 통합관제센터의 경우 1인당 평균 605대를 관제하고 있어, 행정안전부 기준 1인당 적정 CCTV 관제대수 50대에 비해 월등히 상회하고 있다[1].

이로 인해 관제 요원의 피로 누적과 함께 정확도가 떨어지는 문제가 발생한다. 연구에 따르면 22분 이상 지속하여 관제할 경우 95% 까지 관제 효율이 떨어질 수 있어, 최근 영상 관제 시스템은 단순히 영상 송수신 저장에 그치지 않고 사람이나 차량을 식별하고 폭행, 쓰러짐, 침입과 같은 이상 행위를 탐지하는 기술들을 도입하여 지능형 영상 관제 시스템으로 발전하고 있다[2]. 하지만, 다수의 사람들이 존재하는 CCTV 특성으로 기존의 영상 기반 동작 인식 기술은 타인으로 인한 가려짐으로 인식이 어렵고, 학습 모델이 무거운 단점이 있다[3, 4].

이러한 문제를 해결하기 위해, 본 논문에서는 가려짐에 강인하고, 상대적으로 적은 연산으로 동작이 가능하도록 1차원 합성곱(1D Convolution)과 GRU(Gated Recurrent Unit) 기반의 자세 기반 위협 동작 인식 기술을 제안한다. 1차원 합성곱으로 추출한 공간 특징으로 가려짐에 강인한 자세 데이터를 구성하고, GRU의 시간 연속성을 활용하여 위협 동작을 인식하였다. 제안한 모델은 Throw, Protest, Stand, Walk, Run 클래스를 평균 0.4ms의 추론 시간으로 86.84%의 정확도를 달성하였으며, 관제 영상에서 위협 동작을 효율적으로 탐지할 수 있는 가능성을 제시하였다.

Ⅱ. 본론

본 연구는 위협 동작 인식을 위한 데이터세트 구축하고, 실시간으로 동작하는 자세 기반 위협 동작 인식 모델을 개발하였다. 영상 관제 상황에서 위협 동작을 인식하기 위해, 일반적인 동작인 Stand, Walk, Run과 위협 동작인 Throw, Protest를 정의하였고, 학습을 위해 공개 데이터[8, 9, 10, 11, 12, 13]와 자체 수집으로 59,375 프레임으로 구성된 데이터세트를 구성하였다. 위협 동작 인식 모델은 입력 영상에서 실시간으로 자세 데이터를 추출하고, 연속된 자세 데이터를 1차원 합성곱 신경망과 GRU로



<그림1. 위협 동작 인식 학습 프레임워크>

구성된 동작 인식 모델을 통해 실시간으로 위협 동작을 인식할 수 있는 프레임워크를 구현하였다.

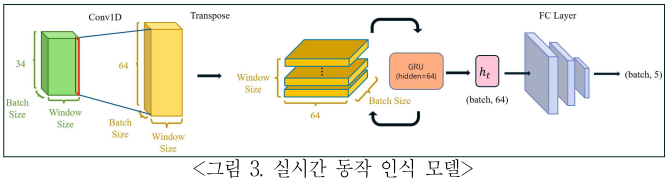
<그림2. 위협 동작 인식을 위한 데이터세트>

기존의 자세 추정 데이터세트는 대부분 정면 영상으로 부터 사용자의 자세를 추정하고 있어, CCTV 영상과 같이 버드아이뷰에서 자세를 레이블링한 데이터는 부족한 상황이다. 이를 위해 CHAD[8], UCF-Crime[9], SCVD[10], ShanghaiTech Campus[11], RWF-2000[12], NTU-RGBD+60[13] 데이터세트에서 영상 데이터를 확보하고 재가공하여 그림 2와 같이 학습용 데이터세트를 구성하였다. 먼저 영상 데이터를 동작 클립별로 분류하고, YOLOX-M[5]으로 사람 객체를 탐지, ByteTrack[6]으로 사람 객체를 구분하고 추적하였으며, HRNEt-w32[7]를 통해 자세 데이터를 추출하는 과정으로 데이터세트를 구성하였다. 영상의 해상도는 다양한 카메라 입력을 고려하여, 320x240 부터, 1920x1080 까지 6종의 해상도로 구성하였다. 또한 데이터세트의 클래스 밸런싱과 보강을 위해 추가로 자체 데이터를 수집 하여, 최종적으로 표1과 같이 데이터 세트를 구성 하였다.

클래스	영상 수	자세 클립 수	전체 프레임	평균 프레임	최소 프레임	최대 프레임
Throw	54	102	8447	98.2	30	173
Protest	75	115	12689	110.3	56	173
Stand	68	121	13241	109.4	46	185
Walk	114	154	17179	111.5	60	193
Run	75	86	8447	98.2	30	173
계	362	578	59375	102.7	30	193

<표 1. 위협 동작 인식을 위한 데이터세트 통계>

† Moonwook Ryu and Joonhyung Kwon contributed equally to this work



<그림 3. 실시간 동작 인식 모델>

그림 3은 제안하는 위협 동작 인식 기술의 실시간 동작 인식 프레임워크이다. 사람 객체 탐지 및 추적 모듈은 데이터세트를 구성할 때 사용하였던, YOLOX-M[5] 과 ByteTrack[6] 으로 구현하였으며, 동작 인식 모델은 그림 3의 실시간 동작 인식 모델과 같이 구현하였다. 본 모델은 연속된 8프레임의 자세 시퀀스로 구성되며, 각 프레임은 K개의 신체 키포인트를 포함한다. 각 키포인트는 D차원의 좌표로 표현되며, 이를 KxD 차원의 벡터로 변환하였다. 입력 데이터 $X \in R^{B\times T\times(K\times D)}$에서 B는 배치 크기를 나타낸다. 키포인트 간의 공간적 관계를 효과적으로 학습하기 위해, 입력 텐서를 $X' \in R^{B\times(K\times D)\times T}$ 형태로 변환한 후, 1차원 합성곱을 적용하였다.

$$F_s = Conv1D(X'; \theta_{conv}) \quad (1)$$

이후 시간적 모델링을 위해 F_s를 다시 $F_s' \in R^{B\times T\times C}$ 형태로 변환하였다. 다음으로 시간적 의존성을 포착하기 위해 GRU[14]를 적용하였다. t 번째 프레임에 대한 GRU의 은닉상태 h_t는 다음과 같이 계산된다.

$$r(t) = \sigma(W_r \cdot F_s'(t) + U_r \cdot h(t-1)) \quad (2)$$

$$z(t) = \sigma(W_z \cdot F_s'(t) + U_z \cdot h(t-1)) \quad (3)$$

$$\tilde{h}(t) = \tanh(W \cdot F_s'(t) + U \cdot (r(t) \odot h(t-1))) \quad (4)$$

$$h(t) = z(t) \odot h(t-1) + (1 - z(t)) \odot \tilde{h}(t) \quad (5)$$

r(t)는 리셋 게이트, z(t)는 업데이트 게이트, h˜(t)는 은닉 상태를 나타내며, ⊙는 요소 별 곱셈을 의미한다. 시공간적 특징이 인코딩 된 마지막 은닉 상태 h(t)는 3개의 완전 연결 레이어를 통해 최종 동작 클래스로 매핑된다

학습 과정에서는 배치 크기 8, 에폭 150, 학습률 3e-4로 설정하였으며, Adam 옵티마이저와 Cross Entropy Loss 손실 함수를 적용하였다. 데이터 전처리 단계에서 윈도우 크기 8, 윈도우 스트라이드 4로 입력 시퀀스를 구성하였고, 클래스 불균형 문제를 해결하기 위해 Weighted Random Sampler를 사용하여 빈도가 낮은 클래스의 샘플이 학습 과정에서 더 자주 등장하도록 조정하였다. 표 2는 각 동작 클래스 및 평균 정확도이다. 전체 데이터세트에서 Top-1 정확도는 86.84%를 기록했으며, 클래스 별로는 Throw 77.78%, Protest 97.96%, Stand 88.64%, Walk 83.33%, Run 83.33%의 정확도를 달성하였다. 또한, 평균 추론 시간은 NVIDIA GeForce RTX 3080에서 0.4ms 확인하였다.

Class Acc.	Throw	Protest	Stand	Walk	Run	Mean
Top-1 Acc.	77.78	97.96	88.64	83.33	83.33	86.84

<표2. 각 동작 클래스 및 평균 정확도>

Ⅲ. 결론

본 연구에서는 지능형 관제 영상 시스템 구현을 위해, 타인으로 인한 가려짐 상황에서 위협 동작을 인식할 수 있는 경량의 학습 모델을 제안하였다. 폭탄 등과 같은 위험한 물체를 던지거나 시위와 같은 이상 행동 및 일반적인 서있기, 걷기, 뛰기 동작들을 인식하기 위한 데이터세트를 구성하였으며, 부족한 데이터는 자체 촬영을 통해 학습 데이터세트를 구축하였다. 모델 측면에서는 자세 데이터를 입력으로 받아 1차원 합성곱으로 공간적 관계를 구성하고, GRU로 시간적 특징을 추출하는 네트워크를 설계하여, 실시간으로 위협 동작을 인식할 수 있는 모델을 구현하였다. 본 학습 모델은 가려짐이 잦은 CCTV 환경에서 자세 정보만으로 위협 행동을 빠르게 탐지 할 수 있어, 과부하된 관제 요원의 인지 부담을 감소시키고, 사고 및 범죄 징후를 조기에 탐지하여 도시 안전 인프라의 효율성과 대응 속도를 크게 향상시킬 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.RS-2024-00462874, 실외 혼잡 환경에서 위협상황 예측 및 선제대응을 위한 센티널 AI (Sentinel AI) 시스템 기술 개발)

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 디지털분야해외석학유치지원 연구결과로 수행되었음 (RS-2024-00459638)

참 고 문 헌

[1] 문성철, 김준철, 이지애, "CCTV에 나타난 범죄 상황 인지 기술 개발", 서울연구원 연구보고서, 2020년 3월

[2] W. Hu, T. Tan, L. Wang, and S. Maybank, “A survey on visual surveillance of object motion and behaviors,”IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 34, no. 3, pp. 334-352, Aug. 2004

[3] R. Poppe, “A survey on vision-based human action recognition,” Image Vis. Comput., vol. 28, no. 6, pp. 976 - 990, Jun. 2010

[4] S. Herath, M. Harandi, and F. Porikli, “Going deeper into action recognition: A survey,” Image and Vision Computing, vol. 60, pp. 4 - 21, 2017

[5] Ge, Zheng, et al. "Yolox: Exceeding yolo series in 2021" arXiv preprint arXiv:2107.08430, 2021

[6] Zhang, Yifu, et al. "Bytetrack: Multi-object tracking by associating every detection box." European conference on computer vision. Cham: Springer Nature Switzerland, 2022

[7] Wang, Jingdong, et al. "Deep high-resolution representation learning for visual recognition." IEEE transactions on pattern analysis and machine intelligence 43.10 (2020): 3349-3364

[8] Danesh Pazho, Armin, et al. "Chad: Charlotte anomaly dataset." Scandinavian Conference on Image Analysis. Cham: Springer Nature Switzerland, 2023

[9] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." Proceedings of the IEEE conference on computer vision and pattern recognition, 2018

[10] Aremu, Toluwani, et al. "SSIVD-Net: A Novel Salient Super Image Classification and Detection Technique for Weaponized Violence." Science and Information Conference. Cham: Springer Nature Switzerland, 2024

[11] Liu, Wen, et al. "Future frame prediction for anomaly detection - a new baseline." Proceedings of the IEEE conference on computer vision and pattern recognition, 2018

[12] Cheng, Ming, Kunjing Cai, and Ming Li. "RWF-2000: An open large scale video database for violence detection." 2020 25th International Conference on Pattern Recognition, 2020

[13] Shahroudy, Amir, et al. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis." Proceedings of the IEEE conference on computer vision and pattern recognition, 2016

[14] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078, 2014