선진국방연구
Journal of Advances in Military Studies https://doi.org/10.37944/jams.v4i1.78
2021, Vol. 4, No. 1, 43-59.

Suspicious behavior recognition using deep learning:

Focusing on keypoint 2D scaling*

Yeonji Park** · Yoojin Jeong*** · Chaebong Sohn****

≪Abstract≫

The purpose of this study is to reinforce the defense and security system by recognizing the behaviors of suspicious person both inside and outside the military using deep learning. Surveillance cameras help detect criminals and people who are acting unusual. However, it is inefficient in that the administrator must monitor all the images transmitted from the camera. It incurs a large cost and is vulnerable to human error. Therefore, in this study, we propose a method to find a person who should be watched carefully only with surveillance camera images. For this purpose, the video data of doubtful behaviors were collected. In addition, after applying a algorithm that generalizes different

* The present Research has been Conducted by the Research Grant of Kwangwoon University in 2020.

** (First Author) Kwangwoon University, Department of Electronics and Communications Engineering, Master's Course Student. kryj9625@kw.ac.kr

*** (Co-Author) Kwangwoon University, Department of Electronics and Communications Engineering, Master Candidate. yoojin2115@kw.ac.kr

**** (Corresponding Author) Kwangwoon University, Department of Electronics and Communications Engineering, Professor, cbsohn@kw.ac.kr

heights and motions for each person in the input images, we trained through a model combining CNN, bidirectional LSTM, and DNN. As a result, the accuracy of the behavior recognition of suspicious behaviors was improved. Therefore, if deep learning is applied to existing surveillance cameras, it is expected that it will be possible to find the dubious person efficiently.

Keywords : defense and security technology, surveillance camera, suspicious person, behavior recognition, OpenPose

Ⅰ. 서론

인구 감소에 따라 거수자 감시 및 경계에 필요한 병력도 감소하고 있어 기존의 인간 중심 경계 감시 체계는 한계가 있다. 이러한 문제를 해결하기 위하여 철조망에 부착된 센서와 CCTV 등으로 구성된 과학화 감시 장비 소요가 증가하고 있으며, 이를 사용하여 외곽이나 철책선, 핵심시설 등 군의 주요 시설 감시 경계를 수행한다. 하지만 늘어나는 CCTV 감시 장비 수에 비해 감시병의 수는 많지 않다. CCTV가 전송하는 화면을 지속해서 모니터링하는 것은 감시병에게 피로감을 유발할 수 있다. 그래서 해당 시스템은 사람의 집중과 판단에 의존하기 때문에 취약점을 가진다. 이러한 단점을 보완하여 실시간으로 영상을 분석하고 위험한 행동을 찾아내어 정보를 알려주는 무인 경계 감시 체계의 중요성이 커지고 있다. 그러나 감시카메라 영상에서 사람의 행동을 정확하게 인식하기 위해서는 몇 가지 어려움이 존재한다. 예를 들어, 화면 안에서 동작을 하는 사람의 위치와 각도 등이 다양하며, 사람에 따라 신장, 비율이 다르고 동작하는 방법이 동일하지 않는 문제가 나타날 수 있다. 마지막으로 낮 시간 외에도 야간에도 감시카메라 영상으로 행동을 인식할 수 있어야 한다.

상기한 문제점을 개선하기 위해 본 연구는 거수자의 행동을 인식하는 방법을 제안한다. 이를 통해 침입, 배회, 싸움, 물건 투척, 포복하는 거수자의 특정 행동 패턴을 탐지하여 경고를 통한 군 시설의 보완 확보와 신속 대응을 목표로 한다. 먼저 영상에서 다양한 각도와 크기로 나타나는 거수자를 인식하기 위해 여러 방향으로 이동하는 거수자 데이터 세트를 제작했다. 그 후 일반화된 동작의 특징 추출을 위한 사람 키포인트의 2D 스케일링 알고리즘이 적용된 딥러닝 기반 행동 인식 모델을 제안한다. 실험 결과 키포인트 2D 스케일링 알고리즘을 사용하였을 때, 정확도가 향상되었다.

Ⅱ. 이론적 배경 및 가설 도출

감시 카메라의 수가 급증하면서 폭력이나 의심스러운 사건을 자동으로 인식하는 시스템의 필요성이 대두되었고 컴퓨터 비전 및 이미지 처리 분야에서 중요한 연구 분야가 되었다. 영상에서 비정상적인 행동의 패턴을 탐지하기 위한 방법은 크게 머신러닝과 딥러닝으로 구분할 수 있다.

머신러닝은 알고리즘을 사용하여 데이터를 분석하고 학습한 후, 그에 따라 결정을 내린다. 머신러닝 알고리즘을 작동하기 위해서 데이터의 특징을 선택 또는 생성하는 Feature engineering 작업이 필요하다. Candamo et al.(2010) 연구에 의하면 대중교통 시스템에서 사고, 범죄, 의심스러운 행동 등을 방지하고 대응하기 위한 행동 인식 기술에서 위험 행동 패턴을 배회(한 사람), 싸움 및 인신공격(여러 사람 간 상호작용), 기물 파손(사람과 차량 간 상호작용), 남겨진 물건 및 침입(사람과 시설의 상호작용)으로 분류하여 방법을 제안하였다. 또한, Park, Song, & Kim(2018) 연구에 의하면 납치 상황에서 특정 관절의 유무, 위치변화, 속도 등의 패턴을 정의하여 분류하였다. Serrano

Gracia et al.(2015)는 모션 블롭을 정의하여 시공간 특징을 추출한 후 분류에 사용하였다. Deniz, Serrano, Bueno, & Kim(2014, January)는 큰 가속도의 존재를 특징으로 SVM을 분류기로 사용하였다. Gao, Liu, Sun, Wang, & Liu (2016)는 Oriented Violent Flows라는 새로운 특징 추출 방법을 제안하였으며 SVM 분류기를 사용하여 폭행 및 비정상적인 행동을 분류하였다. 하지만 Feature engineering을 통해 특징을 정의하여 학습하는 방법은 전문성과 많은 비용이 들고, 특징에 따라 모델 성능에 미치는 영향이 크다는 단점이 있다.

딥러닝은 Feature engineering 없이 연속된 층에서 점진적으로 데이터로부터 의미 있는 표현을 학습하는 방법이다. 대표적인 네트워크는 합성곱 신경망(Convolutional Neural Network, CNN)이다. CNN은 이미지 데이터 처리를 위해 사용되는 네트워크로 이미지 분류, 객체 검출 등에서 좋은 결과를 보여준다(Simonyan & Zisserman, 2014; Szegedy et al., 2015). 더 나아가 이미지들을 묶은 영상에도 CNN을 적용할 수 있다. Redmon et al.(2016)과 Liu et al.(2016)는 실시간 영상에서 물체를 감지하고, 어떤 물체인지 분류가 가능하다. 하지만 이를 사용하여 영상 안에서 연속적으로 발생하고 있는 정보를 알아내기에는 부족하다. 따라서 Ding, Fan, Zhu, Feng, & Jia(2014)는 3d convolution을 사용하여 행동을 인식하였으며, Zhou, Ding, Luo, & Hou(2017)는 움직임 속성을 더 잘 추출하기 위해 RGB 이미지, 광학 흐름 이미지, 가속 이미지의 세 가지 입력으로 학습하는 방식을 사용하였다. 본 연구는 음성, 문자 등 순차적으로 등장하는 데이터 처리에 적합한 순환 신경망(Recurrent Neural Network, RNN) 네트워크를 적용하였다. Zhou, Sun, Liu, & Lau(2015)에서는 CNN의 출력을 LSTM(Hochreiter & Schmidhuber, 1997)에 공급하는 End-to-End 방식의 아키텍처인 C-LSTM 통합 모델을 제안하여 자연어 처리 분야에서 단일 CNN이나 LSTM 모델보다 우수한 결과를 보였다. Sainath, Vinyals, Senior, & Sak(2015)에서는 CLSTM 모델에 추가로 DNN을 결합한 CLDNN 아키텍처를 제안하였으며, 이를 음성인식 분야에 적용하여 향상된 결과를 보였다.

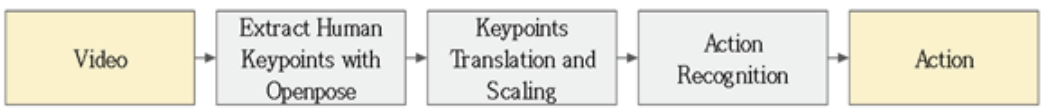
이러한 행동 패턴을 탐지하기 위하여 객체 탐지, 인간 자세 추정 등의 연구가 선행되어야 한다. 예를 들어, Cao, Hidalgo, Simon, Wei, & Sheikh(2019)는 Convolutional Pose Machine(CPM) 기반으로 키포인트를 검출하는 OpenPose 라이브러리를 제안하였으며, Park & Chun(2017)에서는 이를 개선하여 더욱 정확한 다중 신체 키포인트 검출을 위해 RGB-D 정보를 이용하였다. 하지만 RGB-D 기반의 객체 탐지 방식은 객체의 정확한 영역을 제공해야만 키포인트 탐지 정확성을 높일 수 있다. 본 연구에서는 이미지 시계열 데이터인 영상을 처리하기 위해 CLDNN 구조를 개선하였으며, 감시카메라에서 특정 행동 분류에 적용하여 기존 CLDNN 뿐만 아니라 단일 CNN이나 LSTM 및 DNN에서보다 모델의 우수성을 보인다. 또한 추가적인 정보 제공 없이 미리 학습된 네트워크 기반으로 키포인트를 예측할 수 있는 OpenPose를 사용하였다.

Ⅲ. 연구방법

본 논문은 거수자의 특정 행동을 탐지하는 시스템을 제안한다. 제안하는 방법은 다음과 같다.

1. 입력받은 실시간 영상을 OpenPose 라이브러리를 사용하여 사람의 키포인트를 추출한다.
2. 추출한 키포인트에 2D 스케일링 알고리즘을 적용하여 이미지화한 영상을 새롭게 제작한다.
3. 딥러닝 기반 행동 인식 모델에 적용한다.

3.1에서 데이터 수집을, 3.2에서 OpenPose 라이브러리를 사용한 사람 키포인트 추출을, 3.3에서 키포인트 이동 및 2D 스케일링 알고리즘을, 3.4에서 행동 인식을 위한 딥러닝 기반 모델을 설명한다.



<Figure 1> Our suggested system structure

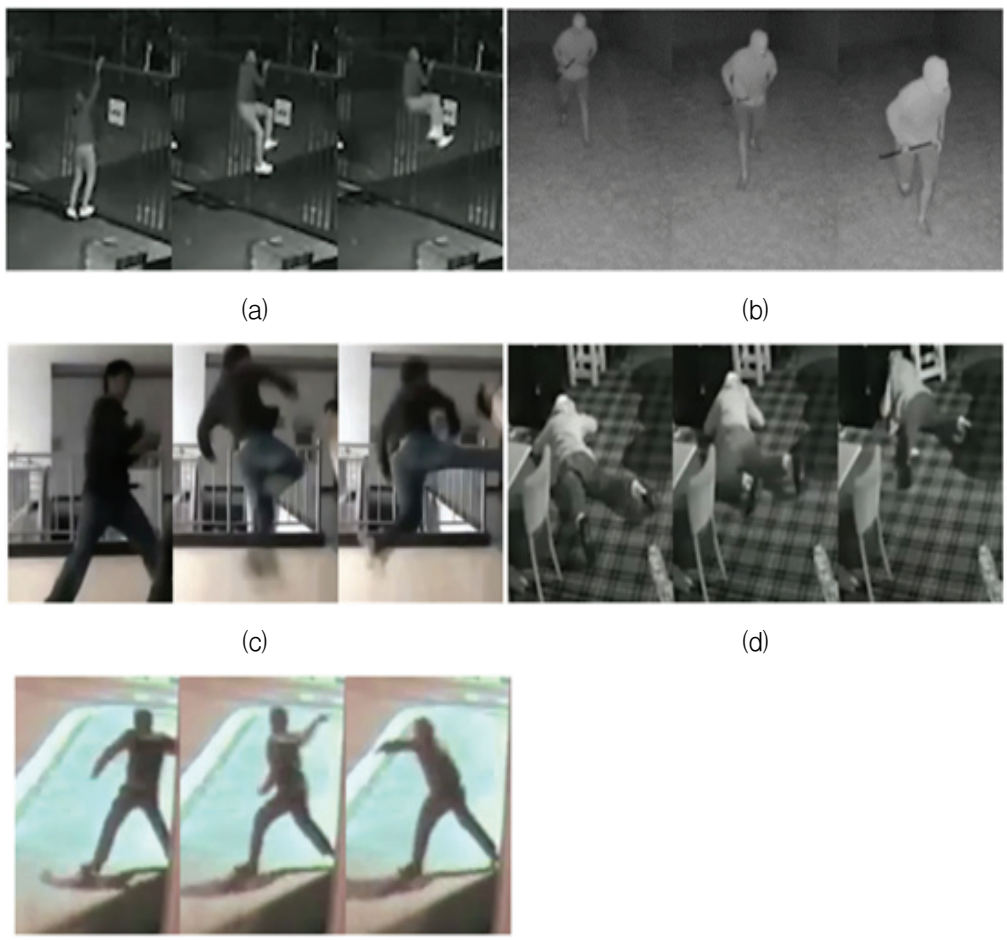
3.1 데이터 수집

거수자 행동 데이터 세트 제작을 위해 1280x720 해상도의 적외선 영상 136개를 수집했다. 데이터 세트는 침입, 배회, 폭행, 포복, 투척의 5개의 클래스로 구성된다. Figure 2는 각 클래스 영상 중 일부 프레임이다.

3.2 OpenPose를 사용한 사람 키포인트 추출

입력 영상에서 주변 환경의 영향을 최소한으로 줄이고, 사람의 행동만을 파악할 수 있도록 사람의 골격과 관절 부위 정보인 키포인트를 추출하여 사용한다. 사람 키포인트 추출을 위해 실시간 단일 이미지에서 사람의 신체, 손, 얼굴 및 발 키포인트를 감지하는 딥러닝 네트워크 라이브러리인 OpenPose를 사용하였다(Cao, Hidalgo, Simon, Wei, & Sheikh, 2019). OpenPose는 BODY_25 모델 기반으로 신체 25개, 왼손 및 오른손 각각 21개, 얼굴 70개로 총 137개의 키포인트를 감지한다.

본 연구에서는 불필요한 데이터를 최소화하기 위해 전체 키포인트 중 신체 키포인트만을 사용한다. Figure 3의 OpenPose 신체 키포인트 중 눈(15, 16), 귀(17, 18), 발가락(19, 20, 22, 23), 발꿈치(21, 24)를 제외한 0번부터 14번까지 15개의 키포인트를 사용했다. Figure 4는 각 클래스 영상에서 사람의 키포인트 추출 예시이다.

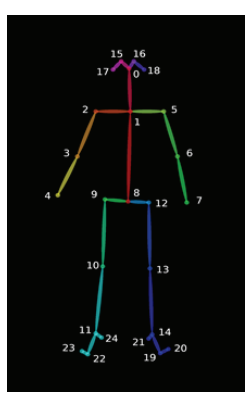


(a) (b)

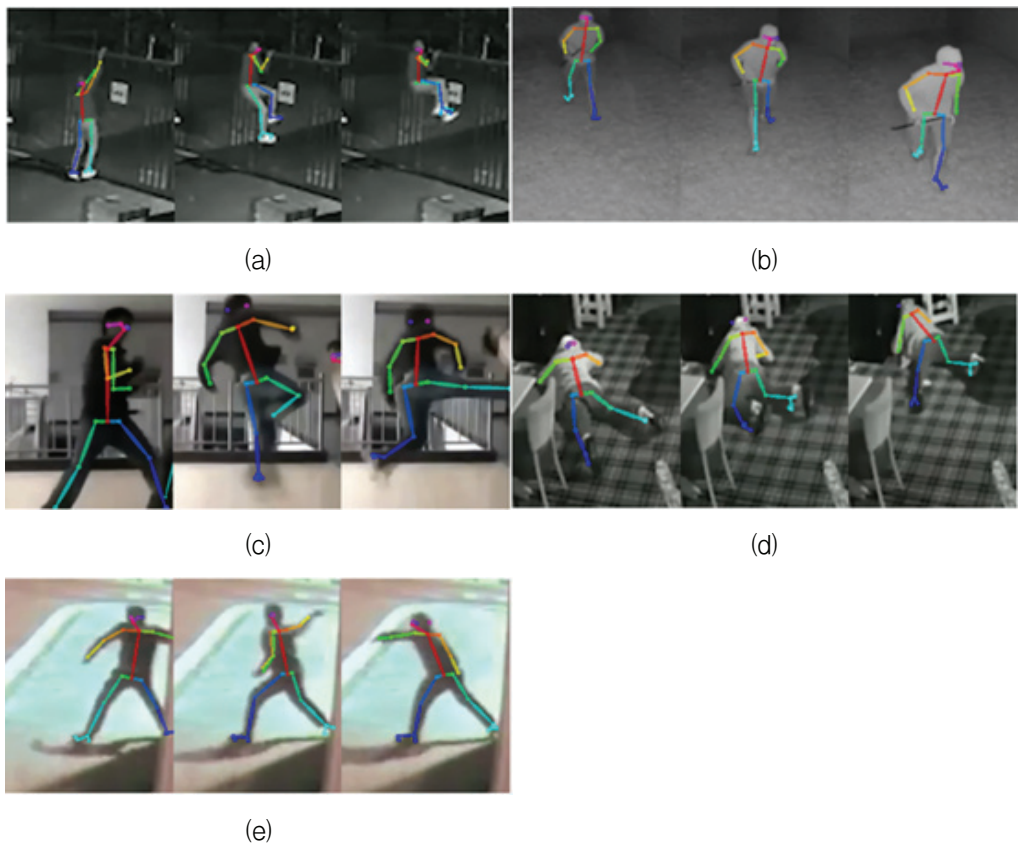
(c) (d)

(e)

<Figure 2> Example frames of five actions in the suspicious-looking people data set. (a) intrusion, (b) wandering, (c) assault, (d) crawling, and (e) throwing objects



<Figure 3> Number assigned to body keypoints in OpenPose



<Figure 4> Example frames of five motions from which human keypoints were extracted. (a) intrusion, (b) wandering, (c) assault, (d) crawling, and (e) throwing objects

3.3 키포인트 이동 및 2D 스케일링 알고리즘

사람마다 신체의 크기와 비율이 다르며 수행하는 동작에도 조금씩 차이가 있다. 이러한 차이를 고려하지 않고 OpenPose를 통해 추출한 키포인트 그대로 사용할 경우 데이터 세트의 특징이 불분명하여 학습이 잘 안될 가능성이 높다. 따라서 영상에서 추출된 신체 키포인트의 상대적인 차이를 제거하기 위해 키포인트 2D 스케일링 작업을 수행한다. 먼저 1280x720 영역 내에 있는 사람의 키포인트의 위치를 각 프레임에서 사람 크기에 해당하는 영역 내로 좌표 이동이 필요하다. 사람의 크기를 구하기 위해 한 프레임에서 키포인트의 x-좌표와 y-좌표를 모은 두 벡터 X와 Y는 식 (1)과 같이 정의한다. 이때 불필요한 프레임을 제거하기 위해 키포인트는 최소 2개 이상인 프레임을 선별하여 사용하였다. 또한 OpenPose에서 인식하지 못한 키포인트의 좌표는 (0, 0)으로 추출되기 때문에 이로 인해 사람 영역의 크기가 실제 크기 값과 달라져 오차가 발생할 수 있다. 이러한 오차를 방지하기 위해 인식되지 않은 키포인트를 제외하여 X와 Y를 계산한다. 여기서, N은 불필요한 키포인트를 제거한 키포인트의 개수이다.

$$X=(x_{0,}x_{1,}\cdots,x_n),\ Y=(y_{0,}y_{1,}\cdots,y_n)\,(1\le n\le N,\ \ 2\le N\le 15)$$

각 X, Y 벡터에서 최댓값과 최솟값의 차를 계산하여 사람의 영역 너비 W와 높이 H를 다음 식 (2)와 같이 구한다.

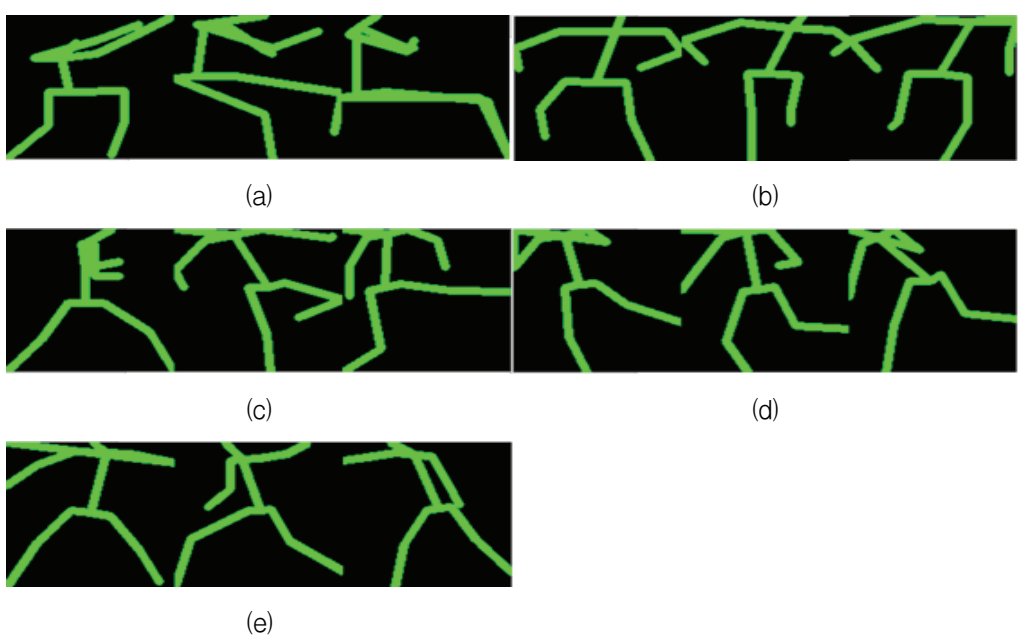
$$W=(X_{\max}-X_{\min}),\ H=(Y_{\max}-Y_{\min})$$

이후 (Xmin, Ymin)의 좌표를 (0, 0)으로 이동시켜 한 프레임 내의 모든 키포인트의 위치를 사람 영역 내로 재조정한다. 재조정된 벡터 X '와 Y'는 다음과 같다.

$$X^{'}=(X_i-X_{\min}),\ Y^{'}=(Y_i-Y_{\min})\,(1\le i\le n)$$

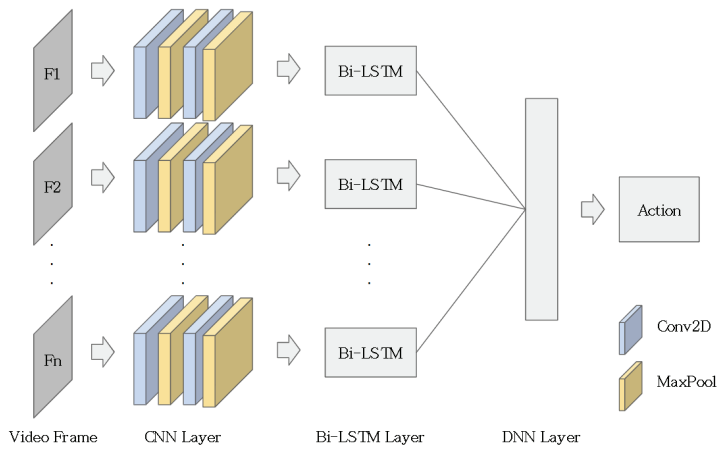
키포인트 재조정이 끝난 후, 128x128의 크기로 스케일링을 수행하여 이미지화한다(식 (4)). X_s와 Y_s는 스케일링된 키포인트 벡터이다. Figure 5는 Figure 4에서 각 클래스에서 추출된 키포인트를 위치 재조정 및 스케일링 적용 후 이미지화한 결과이다.

$$X_s=\frac{X^{'}}{W}*128\ \text{ and }Y_s=\frac{Y^{'}}{H}*128$$



<Figure 5> Example frames imaged after translation and scaling the extracted keypoints. (a) intrusion, (b) wandering, (c) assault, (d) crawling, and (e) throwing objects

3.4 행동 인식 연구



<Figure 6> The modified CLDNN model Flowchart

수정된 CLDNN 모델의 구조는 Figure 6과 같이 CNN과 양방향 LSTM 및 DNN의 세 가지 네트워크로 구성된다. Zhao & Du(2016), Yang et al.(2019)의 연구는 이미지의 공간 정보를 유지하면서 이미지의 특징을 효과적으로 추출하기 위해 Feed-Forward 네트워크인 Convolution Layer의 성능을 입증하였다. 따라서 본 연구에서 전처리한 이미지들을 Convolution Layer에 통과시켜 프레임마다 키포인트의 특징을 추출하였다. Convolution Layer에서는 이미지 데이터에 5x5 필터를 적용한 후 Rectified Linear Unit (ReLU) 활성화 함수와 3x3 필터의 Maxpooling Layer를 추가하여 이를 여러 겹 쌓은 형태로 구성하였다.

여러 연구에서 Convolution Layer 뒤에 LSTM Layer를 추가한 ConvLSTM 네트워크가 시공간 상관관계를 더욱 잘 포착한다는 것을 보였다(Sainath et al., 2015; Shi et al., 2015; Zhou et al., 2015). 연속된 프레임에서 행동의 시계열 특징을 처리하기 위해 ConvLSTM 네트워크 구조를 사용하였다. LSTM은 Feed-Forward 네트워크와 다르게 출력된 결과를 다시 입력으로 넣어 이전의 값과 현재의 값을 같이 고려할 수 있다. 이를 통해 연속된 프레임에서의 키포인트를 사용하여 움직임의 특징을 학습할 수 있다. 하지만 시간 순서대로 입력하기 때문에 정 방향으로만 데이터를 처리하며 현 시점 보다 미래 시점의 데이터는 추론 시 활용할 수 없다. 따라서 Ullah, Ahmad, Muhammad, Sajjad, & Baik(2018)에서 제안한 방식과 같이 역방향으로 처리하는 LSTM을 추가한 양방향 LSTM을 사용하여 성능을 향상시켰다.

마지막으로 양방향 LSTM Layer의 출력을 DNN Layer에 전달한다. DNN Layer에는 256개의 유닛이 존재한다. 이 과정에서 은닉층의 일부 유닛을 임의로 생략하는 Dropout을 추가하여 학습 과정에서 발생할 수 있는 과적합 문제를 해결할 수 있도록 하였다. Dropout을 사용하여 모델을 개

선한 논문이 많이 작성되었으며, Krizhevsky, Sutskever, & Hinton(2017)에서도 AlexNet에 Dropout을 적용하여 성능을 개선한 연구 결과가 있다.

Ⅳ. 실험 및 결과

본 연구에서는 거수자 행동 영상을 데이터 세트로 Table 1과 같이 사용하여 학습을 수행한다. 학습용 데이터 세트에서의 프레임 수는 19,291개, 검증용 데이터 세트에서의 프레임 수는 3,125개, 테스트용 데이터 세트에서의 프레임 수는 8,856개가 사용되었다. 이를 시계열 데이터 학습을 위해 한 영상에서 평균 한 동작 길이인 61 프레임 별로 나눠 하나의 세트로 사용하였고, 61 프레임 미만의 경우 이전 프레임들을 0으로 패딩을 진행하였다. 따라서 학습용 데이터 세트에서의 프레임 묶음 수는 316개, 검증용 데이터 세트에서의 프레임 묶음 수는 51개, 테스트용 데이터 세트에서의 프레임 묶음 수는 145개가 사용되었다.

<Table 1> Overall statistics on the behavioral dataset

Metric	Training	Dev	Test
Number of videos	75	15	46
Number of frames	19,291	3,125	8,856
Number of frame bundles	316	51	145

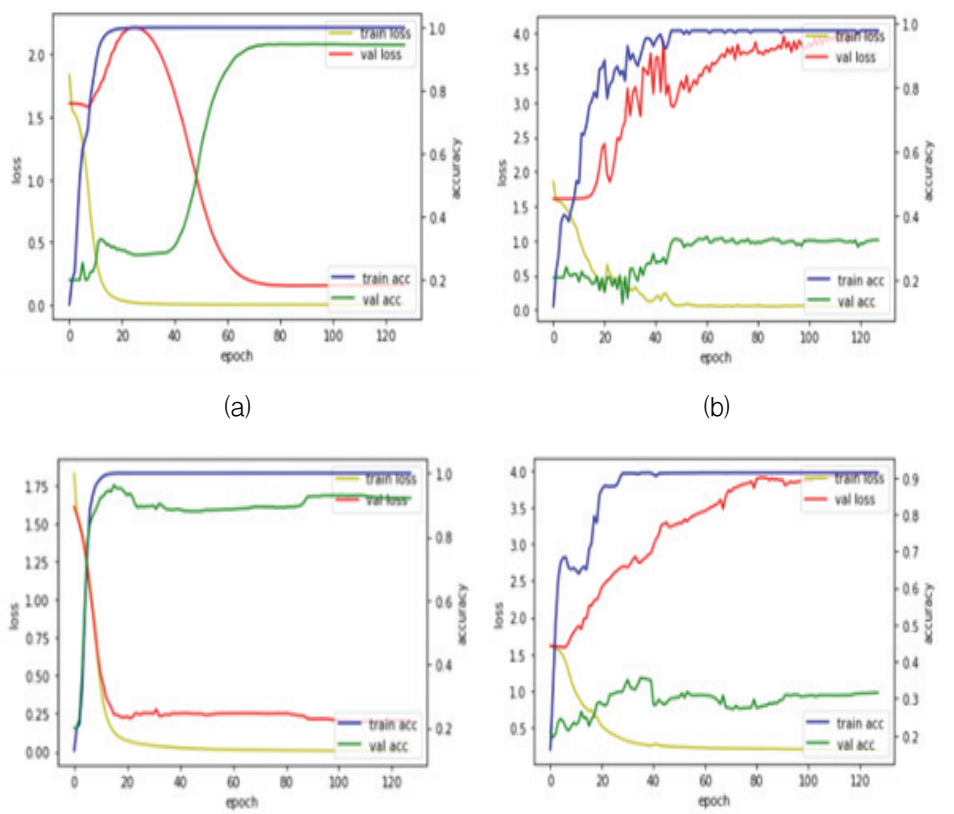
본 실험에서는 키포인트 2D 스케일링을 적용한 사람의 행동 분류 모델을 평가한다. 4.1에서는 단일 모델에 2D 스케일링의 적용 유무에 따른 성능을 비교한다. 4.2에서는 Sainath, Vinyals, Senior, & Sak(2015)에서 제안한 결합 모델인 CLDNN과 다른 RNN 계열 네트워크를 사용한 결합 모델과의 성능 비교를 수행했다. 또한 2D 스케일링 적용에 따른 성능도 비교하였다. 결합 모델에서는 6개의 CNN Layer, 128개의 유닛을 가진 RNN 계열의 Layer와 64개의 완전 연결된 은닉층을 가진 DNN layer로 구성되었다. Dropout은 [0.25, 0.5]에서 선택되었으며, optimizer로는 Adam을 사용하였다. 학습 속도는 1e-3, 배치 사이즈는 16으로 설정하였다. 공정성을 위해 키포인트 2D 스케일링 알고리즘 적용 이외에 다른 설정은 변경 없이 성능 비교를 수행하였다.

4.1 2D 스케일링을 적용한 단일 모델 평가

Table 2에서는 사람의 행동 분류를 위해 단일 CNN 및 LSTM 모델의 성능을 평가한 결과이다. 단일 모델에 2D 스케일을 적용한 결과 정확도가 3배 이상 증가했다. Figure 7은 해당 실험 그래프이다.

<Table 2> Results of applying 2D scaling to a single model

Model	Test accuracy with keypoint 2D scaling (%)	Test accuracy without keypoint 2D scaling (%)
CNN	86.5	28.5
LSTM	86.7	23.9



<Figure 7> The graphs demonstrate the effect of 2d scaling in the single models. Each graph shows the training accuracy, validation accuracy, training loss and validation loss. The items are as follows: (a) CNN model result with keypoints 2D scaling; (b) CNN model result without keypoints 2D scaling; (c) LSTM model result with keypoints 2D scaling; (d) LSTM model result without keypoints 2D scaling

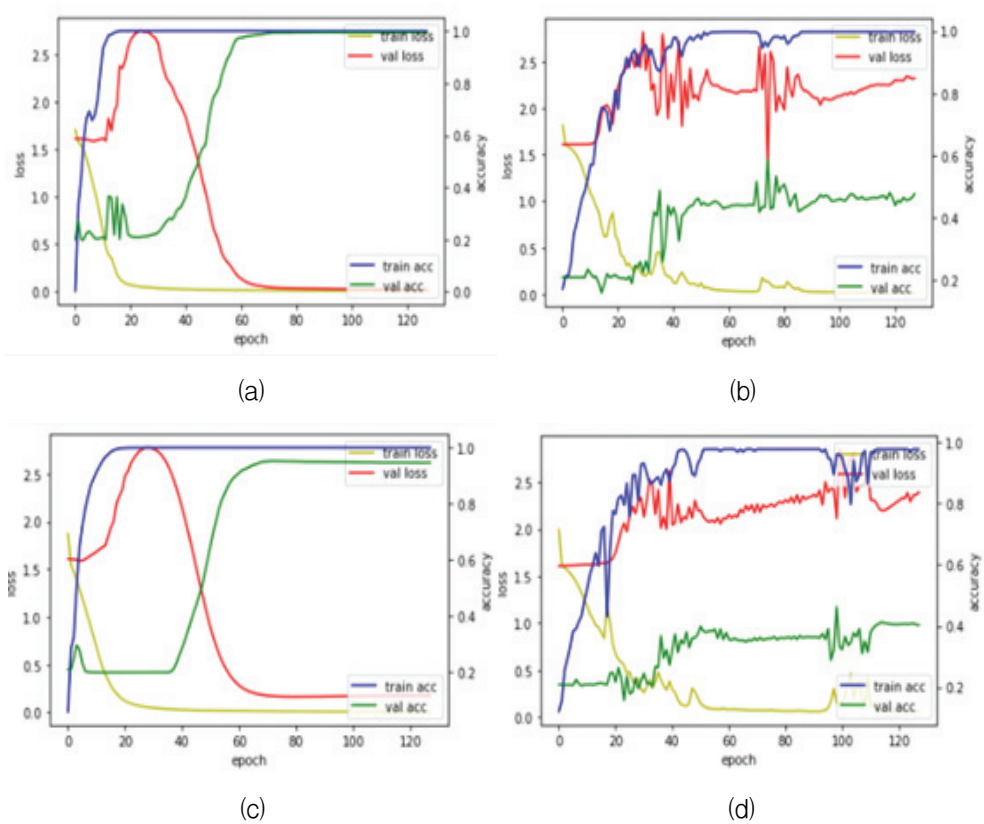
4.2 2D 스케일링을 적용한 결합 모델 평가

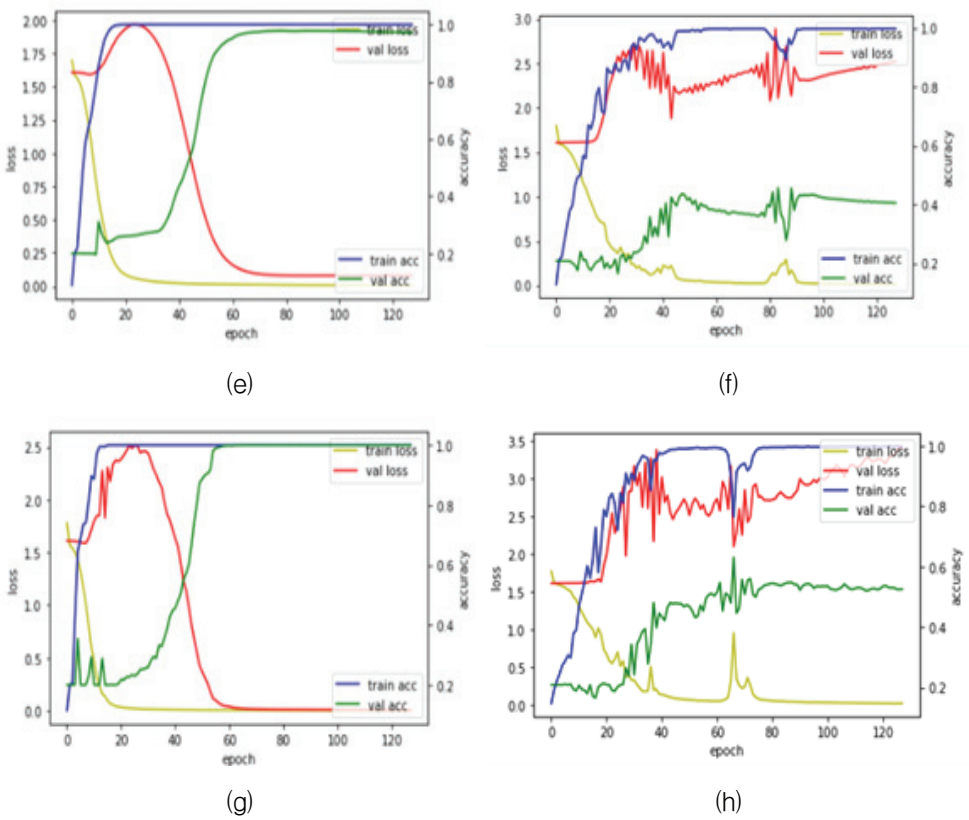
기존의 CLDNN과 CLDNN의 LSTM 레이어를 다른 RNN 계열의 네트워크인 단순 RNN, GRU,

양방향 LSTM 로 변경한 모델들의 성능 비교를 수행하였다. 각 모델의 유닛 수는 128로 동일하며, 단일 레이어로 진행하였다. Table 3은 결합 모델들의 성능 평가 결과를 보여주며, Figure 8은 해당 실험 그래프이다. 모든 모델에 키포인트 2D 스케일링을 적용 유무를 비교했을 때, 2D 스케일링을 적용한 경우 성능이 크게 향상되었음을 볼 수 있다. 키포인트 2D 스케일링을 적용한 경우 성능이 약 2.8배 정도 향상되었다. 또한, Table 2의 단일 모델과 비교를 통해 결합 모델의 성능이 더 좋음을 알 수 있으며, 그중 GRU나 양방향 LSTM를 적용했을 때, 기존 CLDNN 보다 더 나은 결과를 보여주었다.

<Table 3> Results of applying 2D scaling to a combined model

Model	Test accuracy with keypoint 2D scaling (%)	Test accuracy without keypoint 2D scaling (%)
CLDNN	77.6	27.1
CNN+RNN+DNN	44.0	23.0
CNN+GRU+DNN	90.9	33.5
CNN+Bi-LSTM+DNN	92.6	32.0





<Figure 8> The graphs demonstrate the effect of 2d scaling in the combined models. Each graph shows the training accuracy, validation accuracy, training loss and validation loss. The items are as follows: (a) CLDNN model result with keypoints 2D scaling; (b) CLDNN model result without keypoints 2D scaling; (c) CNN+RNN+DNN model result with keypoints 2D scaling; (d) CNN+RNN+DNN model result without keypoints 2D scaling; (e) CNN+GRU+DNN model result with keypoints 2D scaling; (f) CNN+GRU+DNN model result without keypoints 2D scaling; (g) CNN+Bi-LSTM+DNN model results with keypoints 2D scaling; (h) CNN+Bi-LSTM+DNN model results without keypoints 2D scaling

V. 결론

본 논문은 감시 카메라 영상에서 거수자의 행동을 인식하는 방법을 제안한다. 적외선 영상에서 사람 키포인트를 추출한 후, 2D 스케일링을 적용함으로써 감시 카메라 영상에서 사람마다 다른 특징들을 일반화하여 정확도를 높였다. 그리고 본 연구의 실험 결과를 통해 양방향 LSTM을 사용했을 때 더 우수한 성능으로 행동을 분류할 수 있음을 알 수 있었다. 하지만 데이터 세트의 부족으로 실제 상황에서 발생하는 다양한 위험 행동을 인식하는 데 부족함이 있어 영상에 많은 사람이 나타

날 경우에 실제 행동을 인식하는 데 어려움이 있다. 그러므로 향후 연구가 적용되기 위해서 위험 상황에서 여러 사람 간, 사람과 사물 및 사람과 시설 간의 상호작용에 관한 추가적인 연구가 필요하다. 이를 위해 후속연구는 실시간 시스템 및 최신 네트워크에 대한 연구와 적용이 지속될 필요가 있다. 이러한 한계점을 보완하여 감시 카메라에 적용한다면 거수자를 찾는데 효율적인 보조 시스템이 될 것이며, 군 내외 주요 시설의 보안 강화에 도움이 될 것이다.

Acknowledgements

We would like to thank Editage (www.editage.co.kr) for English language editing.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Reference

Candamo, J., Shreve, M., Goldgof, D. B., Sapper, D. B., & Kasturi, R. (2010). Understanding Transit Scenes: A Survey on Human Behavior-Recognition Algorithms. *IEEE Transactions on Intelligent Transportation Systems, 11*(1), 206-224. https://doi.org/10.1109/tits.2009.2030963

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(1), 172-186. https://doi.org/10.1109/TPAMI.2019.2929257

Deniz, O., Serrano, I., Bueno, G., & Kim, T. K. (2014, January). Fast violence detection in video. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP), 2*, 478-485. IEEE. https://ieeexplore.ieee.org/document/7294968

Ding, C., Fan, S., Zhu, M., Feng, W., & Jia, B. (2014, December). Violence detection in video by using 3D convolutional neural networks. In *International Symposium on Visual Computing* (pp. 551-558). Springer, Cham. https://doi.org/10.1007/978-3-319-14364-4_53

Gao, Y., Liu, H., Sun, X., Wang, C., & Liu, Y. (2016). Violence detection using Oriented VIolent Flows. *Image and Vision Computing, 48-49*, 37-41. https://doi.org/10.1016/j.imavis.2016.01.006

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84-90. https://doi.org/10.1145/3065386

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2

Park, J. H., Song, K., & Kim, Y. S. (2018). A kidnapping detection using human pose estimation in intelligent video surveillance systems. *Journal of the Korea Society of Computer and Information, 23*(8), 9-16. https://doi.org/10.9708/jksci.2018.23.08.009

Park, S., & Chun, J. (2017). A Method for Body Keypoint Localization based on Object Detection using the RGB-D information. *Journal of Internet Computing and Services, 18*(6), 85-92. https://doi.org/10.7472/jksii.2017.18.6.85

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779-788. https://doi.org/10.1109/cvpr.2016.91

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4580-4584). IEEE. https://doi.org/10.1109/icassp.2015.7178838

Serrano Gracia, I., Deniz Suarez, O., Bueno Garcia, G., & Kim, T.-K. (2015). Fast Fight Detection. *Plos ONE, 10*(4), e0120448. https://doi.org/10.1371/journal.pone.0120448

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems, 28*, 802-810. https://dl.acm.org/doi/10.5555/2969239.2969329

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint*. https://arxiv.org/abs/1409.1556

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D, & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1-9. https://doi.org/10.1109/cvpr.2015.7298594

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., & Baik, S. W. (2018). Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. *IEEE Access, 6*, 1155-1166. https://doi.org/10.1109/access.2017.2778011

Yang, A., Yang, X., Wu, W., Liu, H., & Zhuansun, Y. (2019). Research on Feature Extraction of Tumor Image Based on Convolutional Neural Network. *IEEE Access, 7*, 24204-24213. https://doi.org/10.1109/access.2019.2897131

Zhao, W., & Du, S. (2016). Spectral‐Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach. *IEEE Transactions on Geoscience and Remote Sensing, 54*(8), 4544-4554. https://doi.org/10.1109/tgrs.2016.2543748

Zhou, C., Sun, C., Liu, Z., & Lau, F. (2015). A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630*

Zhou, P., Ding, Q., Luo, H., & Hou, X. (2017, June). Violent interaction detection in video based on deep learning. *Journal of Physics: conference series, 844*(1), 012044. https://doi.org/10.1088/1742-6596/844/1/012044

원 고 접 수 일 2020년 11월 09일
원 고 수 정 일 2021년 03월 04일
게 재 확 정 일 2021년 04월 23일

국문초록

https://doi.org/10.37944/jams.v4i1.78

딥러닝을 이용한 군 내외 거수자 행동 인식:
키포인트 2D 스케일링을 중심으로*

박연지** · 정유진*** · 손채봉****

본 연구는 딥러닝을 통해 군 내외에서 거수자의 행동을 인식하여 국방 보안 체계를 강화하는 데 목적을 두고 있다. 감시 카메라는 범죄자, 수상한 행동을 하는 사람을 찾아내는데 도움이 된다. 하지만 관리자가 카메라에서 전송되는 많은 영상을 모두 모니터링해야 한다는 점에서 비효율적이다. 큰 비용이 발생하며, 인적 오류에 취약하다. 따라서 본 연구에서는 감시카메라 영상만으로 주의 있게 봐야 할 행동을 하는 사람을 찾아내는 방법을 제안한다. 이를 위해 거수자 영상 데이터를 수집하였다. 또한, 입력 영상에서 사람마다 다른 신장, 동작을 일반화하는 알고리즘을 적용한 후 CNN, 양방향 LSTM, DNN을 결합한 모델을 통해 학습하였다. 실험 결과, 거수자의 행동 인식의 정확도를 향상시켰다. 따라서 기존 감시 카메라에 딥러닝을 접목한다면 거수자를 효율적으로 찾아낼 수 있을 것으로 기대한다.

주제어 : 국방보안기술, 감시카메라, 거수자, 행동 인식, 오픈포즈

* 이 논문은 2020년도 광운대학교 융·복합 연구과제 지원사업의 지원을 받아 수행된 연구임.
** (제1저자) 광운대학교 전자통신공학과, 석사과정, kryj9625@kw.ac.kr
*** (공동저자) 광운대학교 전자통신공학과, 석사수료, yoojin2115@kw.ac.kr
**** (교신저자) 광운대학교 전자통신공학과, 교수, cbsohn@kw.ac.kr