CCTV 원본 영상과 추출된 스켈레톤 영상을 함께 이용하는 폭력 인식기

주현성, 김유성
인하대학교 정보통신공학과
altair2881@gmail.com, yskim@inha.ac.kr

Violence detector using both CCTV videos and extracted skeleton images

Hyun-Seong Joo, Yoo-Sung Kim
Dept. of Information and Communication Engineering, Inha University

요 약

본 논문은 영상 속 폭력행위를 인식하기 위해 3 차원 컨벌루션을 활용하여 원본 영상과 스켈레톤(skeleton)영상으로부터 추출한 시각 및 움직임 정보를 동시에 활용하는 2-스트림 구조의 폭력상황 인식기를 제안한다. 제안된 폭력상황 인식기에서는 수평, 수직 방향의 큰 움직임이 많이 나타나는 폭력영상의 특성을 활용하기위해 각 방향의 특성을 독립적으로 학습할 수 있는 split-FAST 3 차원 컨벌루션을 활용하고, 3 차원 Attention 을 적용하여 시각 및 움직임 정보 추출 시 영상의 중요지역을 중점적으로 반영하도록 함으로써 촬영 기기의 이동 또는 여러 사람의 뒤엉킴 등으로 영상의 시점 변화나 상황 변화가 잦은 경우에도 강인한 성능을 가질 수 있도록 하였다. 또한 기존의 연구들과 달리 비제약적인 환경에서 CCTV, 모바일 카메라 등으로 촬영된 실제 영상들로 구성된 RLVS 데이터셋을 학습 데이터로 사용함으로써 실제의 폭력 행위를 잘 인식할 수 있도록 하였다. RLVS 를 이용한 평가 실험에서 제안된 폭력상황 인식기가 약 92%의 인식 정확도를 얻었다.

1. 서론

지능형 CCTV 를 위한 위험상황 및 사람의 행위 인식의 성능을 개선하기 위해 딥러닝 모델을 사용하는 연구 동향에 따라 CCTV 영상으로부터 폭력행위를 인식하기 위해서 딥러닝을 적용하려는 연구가 많이 시도되고 있다[1-3]. 과거 폭력행위 인식 연구들은 영상 데이터의 원본을 활용해 추출한 시각적 특징[1]이나, Optical-flow[2] 또는 연속된 프레임 간의 차분 영상[3] 등과 같이 움직임과 관련된 정보로부터 추출한 특징들을 활용하여 폭력 행위를 구분하였다[1-3]. 그러나 실제환경에서 취득한 폭력영상[1]의 경우를 살펴보면 통제되지 않은 환경에서 다수의 출연자가 동시다발적으로 일으키는 다양한 폭력행위들이 연속적으로 나타나고, 격렬한 몸싸움이나 촬영자의 물리적 이동 등으로 인해 화면의 빠른 전환 또는 촬영의 시점(viewpoint) 변화가 복잡하게 이루어지는 특성을 가지고 있다. 따라서 실제 폭력행위를 정확하게 검출하기 위해서는 움직임 정보를 충실하게 표현할 수 있는 특징들을 효율적으로 추출하여 사용하고 화면전환이나 시점 변화에 강인한 폭력상황 인식기가 필요하다.

본 연구에서는 OpenPose 라이브러리[4]를 활용해 영상에 나타난 사람의 움직임 정보만을 담은 스켈레톤 영상을 추출하고, 이를 원본 영상과 함께 모델의 입력 데이터로 활용해 폭력행위를 인식하는 2-스트림 구조의 3 차원 컨벌루션(3-Dimention convolution) 기반 폭력행위 인식모델을 제안한다. 이때 영상에 나타나는 종적, 횡적 특성을 보다 분별력 있게 추출하기 위해 일반적인 3 차원 컨벌루션이 아닌 split-FAST(Fractioned Adjacent Spatial and Temporal) 3 차원 컨벌루션[5]을 사용하며, 영상의 중요부분만을 학습에 반영하여 장면의 잦은 전환과 시점변화에 강인한 모델을 만들기 위해 공간중심 어텐션(Spatial Attention)[6] 기법을 적용하였다.

2. 관련연구

폭력행위를 인식하기 위한 기존의 연구[1-3]들은 AlexNet, VGG-16 과 같은 2 차원 컨볼루션 인공신경망을 활용하여 동영상의 각 프레임 이미지로부터 공간적 특징을 추출하고 LSTM, ConvLSTM 과 같은 순환 인공신경망을 활용하여 추출된 공간적 특징들간의 시계열 특징을 추출하여 활용하였다. [1]에서는 원본 영

상으로부터 추출한 시각적 특징만을 활용하고, [2,3]은 연속된 프레임 사이의 x, y 방향 움직임을 계산하는 Optical-flow 나 연속된 프레임 사이의 단순 차분 연산과 같은 방법으로 생성한 움직임 특징만을 활용한다는 차이가 있다.

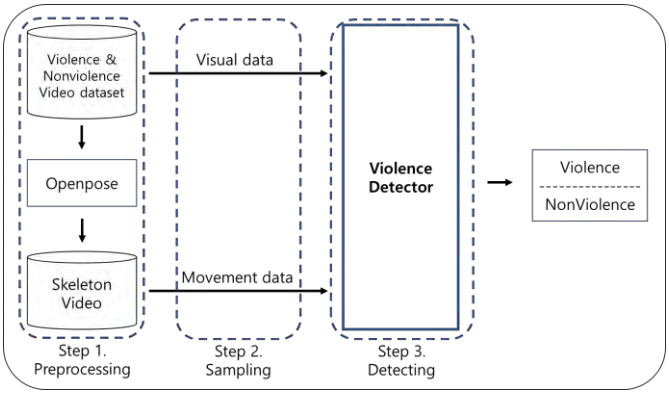
하지만 앞서 언급한 바와 같이 실제 환경에서 취득한 폭력행위 영상들은 일반 영상들과 다른 특징들을 가지고 있다. 따라서 [1]과 같이 원본 영상만을 사용할 경우 폭력행위 인식에 중요한 요소가 될 수 있는 움직임 정보를 모델에 반영하기 어렵고, [2,3]과 같이 비교적 단순한 방법으로 움직임 정보를 추출할 경우 추출된 특징이 행위자의 움직임에 의한 것인지 카메라의 움직임이나 시점의 변화에 의한 것인지 파악하기 어렵다는 문제점이 있다. 이러한 문제들을 해결하기 위한 다양한 방법 중 하나로 [4]에서는 영상에서 나타나는 움직임 요소 중 사람의 움직임 정보를 추출하는 OpenPose 를 제안하였다. OpenPose 는 해부학적으로 중요한 사람의 25 개 관절의 영상 내 위치정보와 그들 사이의 연결관계를 영상에 나타낸 스켈레톤(Skeleton)이미지를 각 사람마다 실시간으로 예측한다. OpenPose 를 통해 추출한 스켈레톤 이미지를 영상 속 출연자의 움직임 정보로 활용하여 위의 문제들을 해결하고자 한다.

최근 행위인식 분야에서는 기울기 소멸(Gradient vanishing)의 잦은 발생과 같은 2 차원 컨볼루션 인공신경망과 순환 인공 신경망이 결합된 구조의 단점을 회피하고 영상의 시공간적 정보를 추출하는 방법으로써 3 차원 컨벌루션 기반의 모델을 활용하는 연구가 늘고 있다[5]. 그 중 split-FAST 3 차원 컨벌루션은 영상에 나타나는 종적, 횡적 특성을 보다 분별력 있게 추출하고자 기존의 3 차원 컨벌루션 연산을 시간, 공간에 대한 각각의 연산으로 분리하여 먼저 계산한 후 이를 결합한다.

또한 폭력행위 영상의 또 다른 특징인 잦은 화면전환 및 시점의 변화 등에 의한 잡음(noise)은 학습 단계에서 모델이 빠르게 수렴하는 것을 방해하고 실제 활용 단계에서는 오분류를 일으키는 원인으로 작용할 수 있다. 이러한 문제를 해결하기 위해 영상의 영역 중 폭력행위의 판단에 중요한 시공간적 영역을 구분하여 활용할 수 있는 어텐션 기법(Attention mechanism)[6]을 활용한다. 특히 공간중심 어텐션의 경우 입력 영상의 중요한 영역을 중심으로 학습과정이 수행되도록 함으로써 영상에 나타나는 잡음에 의한 부정적 영향을 최소화할 수 있다.

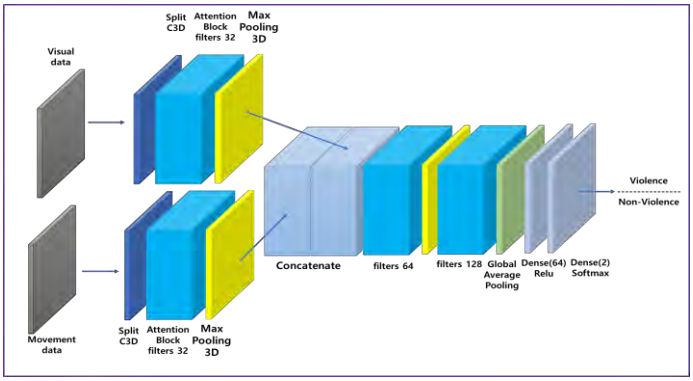
3. 본론

제안한 폭력 인식기의 처리 과정은 [그림 1]과 같이 세 개의 주요 단계로 구성된다. 첫 번째 단계인 전처리과정에서는 Openpose 라이브러리[4]를 활용하여 입력 영상 속의 사람 움직임에 관한 정보만을 추출하여 스켈레톤 영상으로 변환한다. 인접한 프레임 사이에는 중복도가 높기 때문에 영상의 모든 프레임을 사용하면 불필요한 계산량의 증가를 초래하게 되므로, 이어지는 두 번째 단계에서 앞서 생성한 스켈레톤 영상과 원본 영상들로부터 각각 t 개의 프레임을 샘플링한다. 샘플링 과정은 영상의 전체 프레임을 t 로 나누어 샘플링 간격을 설정하고 영상의 첫 프레임으로부터 해당 간격마다 프레임을 추출하도록 하였다. 이때 t 는 실험을 통해 30 으로 설정하였다. 만들어진 원본 영상의 시각 데이터(Visual data)와 스켈레톤 영상의 움직임 데이터(Movement data)들은 세 번째 단계인 폭력행위 인식 모델(Violence Detector)의 입력으로 함께 전달되어 해당 영상의 폭력행위 여부를 예측하는데 활용된다.



(그림 1) 전체적인 구조

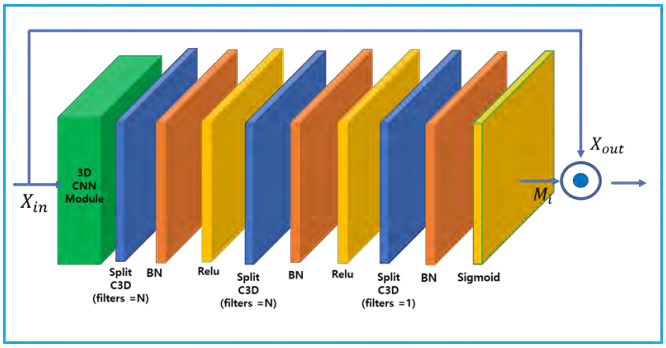
[그림 2]와 같이 폭력 행위 인식 모델은 앞서 만든 시각 데이터와 움직임 데이터를 동시에 입력으로 받을 수 있도록 2-스트림 구조를 가지며 3 차원 컨벌루션을 통해 각각의 입력으로부터 추출해 낸 시공간적 특징들을 융합하여 해당 영상의 폭력행위 포함 여부를 판단하는데 사용한다.



(그림 2) 폭력 행위 인식 모델 (Violence Detector)

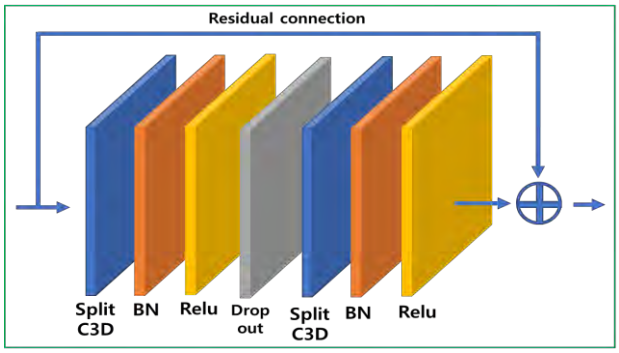
폭력상황 인식기를 구성하는 컨벌루션 블록의 수, 각 계층의 출력 차원 수 등과 같은 네트워크의 상세 구조는 스켈레톤 데이터로부터 분별력 있는 특징을 추출하는데 우수한 성능을 가진 것으로 알려진 AGC-Net 의 구조를 참고하여 재구성하였고 두개의 특징은 네트워크 중간 단계에서 융합하는 것으로 하였으며 중간의 어느 계층에서 융합하는 것이 좋은지는 실험을 통해 결정하였다.

입력으로 들어온 각각의 시각, 움직임 프레임 세트는 먼저 3 차원 컨벌루션을 거쳐 시공간적 특징맵으로 변환된다. 이때 사용된 3 차원 컨벌루션은 폭력영상에 자주 나타나는 수직, 수평 방향의 큰 움직임에 관한 특징들이 잘 인지될 수 있도록 split-FAST 3 차원 컨벌루션을 사용하였다. 이어지는 어텐션 블록에서는 입력(X_{in})으로 들어온 특징 맵 중 결과의 판단에 중요한 부분만을 추려내는 작업을 수행하기 위해 [그림 3]과 같이 입력의 중요 부분을 나타내는 가중치 맵(M_i)과 입력(X_{in})을 대응원소 곱(element-wise multiplication)으로 계산한 X_{out}을 출력한다. 가중치 맵(M_i)를 학습하기 위한 어텐션 블록의 구조는 3 차원 인공신경망 모듈 및 각각 3 개씩의 3 차원 컨벌루션, Batch Normalization, Activation 계층으로 이루어져 있으며 마지막 컨벌루션의 출력 차원을 1 로, 활성화(activation) 함수를 시그모이드(sigmoid)로 설정하여 입력(X_{in})의 중요 부분을 확률적으로 예측할 수 있도록 하였다.



(그림 3) 어텐션 블록 (Attention block)

이 블록의 전체 구조는 [6]의 공간중심 어텐션 구조를 참고하였으나 2 차원 컨벌루션을 사용하여 공간적인 중요부분만 인식하던 [6]과 달리 3 차원 컨벌루션을 사용하여 공간적인 중요부분 뿐 아니라 시간적인 중요부분도 인식할 수 있도록 하였다. 또한 3 차원 인공신경망 모듈(3D CNN Module)은 입력 특징맵이 가진 움직임적 요소를 분별력 있게 추출할 수 있도록 하기 위해 [그림 4] 와 같이 [7]의 AGC- Block 과 동일한 구조를 사용하였다.



(그림 4) 3 차원 인공신경망 모듈 (3D CNN Module)

이처럼 어텐션을 통해 중요 부분이 강조된 두 특징맵들은 (2*2*2) 크기의 커널을 사용하는 Max Pooling 계층을 통해 크기가 반으로 줄어든 후 결합(concatenate)과정으로 서로 이어 붙여진다. 이렇게 하나가 된 특징맵은 두 번의 어텐션 블록과 Max pooling 을 더 거친 후 GAP 를 통해 특징 벡터로 변환되며, 이때 각 블록의 필터 수는 [7]의 구조와 같이 블록을 거칠 때 마다 2 배씩 늘려가도록 하였다. 이렇게 만들어진 특징 벡터는 두 개의 연속된 Dense 계층으로 이루어진 classifier 가 해당 영상이 폭력행위를 포함하고 있는지를 판단하기 위해 사용된다.

4. 실험 및 평가

본 연구에서는 모델의 학습에 있어 보다 실제적이고 비제약적인 환경에서 취득한 데이터를 활용하기 위해 [1]에서 제안한 “Real-Life Violence Situations” (RLVS) 데이터셋에 포함된 1000 개의 폭력 영상과 1000 개의 비폭력 영상을 사용하였다.

먼저 첫번째 실험에서는 움직임 정보를 얻기 위해 사용된 스켈레톤 데이터가 폭력행위 인식에 있어 유익한 효과를 줄 수 있는지 그리고 어떻게 사용하는 것이 효과적인지 확인하기 위해 비교해 보았다.

<표 1> 스켈레톤 영상의 사용방법에 따른 비교

입력 방법	정확도	
	평균	최대
원본 영상만 입력	0.727	0.750
원본 영상에 스켈레톤을 오버랩시킨 단일 영상 입력	0.757	0.791
원본 영상과 스켈레톤 영상을 분리 입력	**0.917**	**0.927**

실험 결과에 따르면 본 연구에서 제안한 방법과 같이 원본 영상과 스켈레톤 영상을 구분하여 2-스트림으로 입력하는 경우가 원본 영상만을 입력하는 경우와 두 영상을 통합하여 단일 스트림으로 입력하는 경우보다 평균적으로 약 16% 더 나은 성능을 보였다. 이는 폭력행위 인식에 있어서 움직임 정보인 스켈레

톤 정보를 시각적 정보와 함께 활용하는 것이 더 효과적이며 스켈레톤 정보를 사용할 때는 원본 영상과 독립적인 별도의 입력으로 사용하는 것이 더 효과적이라는 의미로 볼 수 있다.

두번째 실험에서는 split-FAST 3 차원 컨벌루션을 사용하는 것이 폭력영상을 구분하는데 효과가 있는지 측정하기 위해 일반적인 3 차원 컨벌루션인 C3D 를 사용한 경우와 split-FAST C3D 를 사용한 경우를 서로 비교하였다. 이때 사용한 네트워크는 입력영상을 하나만 수용하며 입력은 원본영상에 스켈레톤 정보를 오버랩시킨 인코딩 영상을 사용하였다.

<표 2> C3D 와 split-FAST C3D 비교

3차원 컨벌루션	정확도	
	평균	최대
기존 C3D	0.727	0.750
split-FAST C3D	**0.755**	**0.786**

실험 결과 split-FAST C3D 를 사용한 경우가 기존의 C3D 를 사용한 경우 보다 폭력 행위 인식에 대해서 평균적으로 약 2.8% 더 나은 성능을 보였다. 이는 수직, 수평방향의 특징을 독립적으로 추출하여 활용하는 것이 폭력행위를 인식하는데 도움을 줄 수 있다는 의미로 볼 수 있다.

세번째 실험에서는 본 논문에서 제안한 딥러닝 모델에서 시각 정보와 움직임 정보로부터 추출한 특징 정보를 어느 레벨에서 융합하는 것이 적절한지를 알아보기 위해 [표 3]와 같이 실험을 진행하였다. 실험 결과에 따르면 어텐션 블록 1 을 지난 특징들을 융합한 경우가 그 외의 경우보다 평균적으로 약 4~12% 더 나은 성능을 보였다.

<표 3> 융합 위치에 따른 비교

입력데이터	융합위치	정확도	
		평균	최대
원본 영상에 스켈레톤을 오버랩시킨 단일 영상 입력	입력 영상	0.735	0.743
원본 영상과 스켈레톤 영상을 분리 입력	**어텐션 블록1**	**0.922**	**0.938**
	어텐션 블록2	0.878	0.901
	어텐션 블록 3	0.801	0.875

각 실험들을 종합하면, 본 연구의 제안처럼 시각 정보와 움직임 정보를 2-스트림으로 입력 받아 처리하고 첫번째 어텐션 블록으로 추출된 특징 벡터를 융합하여 폭력행위를 인식하는 것이 가장 좋은 성능을 낼 수 있음을 확인하였다.

5. 결론

본 논문은 영상에 나타난 폭력 행위를 인식하기 위해 원본 영상으로부터 추출한 시각정보와 스켈레톤 영상으로부터 추출한 움직임 정보를 동시에 활용하는 2-스트림 구조의 3 차원 컨벌루션 기반의 신경망 모델을 제안했다. 특히 수평, 수직 방향의 큰 움직임이 많은 폭력영상의 특성을 모델로부터 추출되는 특징에 반영하기 위해 split-FAST 3 차원 컨벌루션을 사용하고, 영상의 시점이나 흔들림 등의 변화가 많은 폭력영상의 특성상 영상의 중요부분만을 특징에 반영하기 위해 3 차원 어텐션을 적용하였다. 또한 기존의 연구들과 달리 비제약적인 환경에서 취득한 실제 영상들로 구성된 RLVS 데이터셋을 학습 데이터로 사용함으로써 실제의 폭력 행위를 잘 인식할 수 있도록 하였다. RLVS 를 이용한 평가 실험에서 제안된 폭력상황 인식기가 약 92%의 인식 정확도를 얻어 제안 방법이 다른 방법보다 더 나은 성능을 보임을 확인하였다.

향후 연구로는 스켈레톤 영상을 얻기 위해 전처리 단계에서 별도의 라이브러리를 사용하는 번거로움을 없애기 위해 인식 모델의 네트워크 안에서 움직임 정보를 추출하는 방법을 고려하고 시각적 정보와 움직임 정보가 서로 다른 형태를 가지므로 각각에 대한 적합한 특징 추출방법을 연구할 것이다.

참고문헌

[1] Mohamed Mostafa Soliman “Violence Recognition from Videos using Deep Learning Techniques”, Proc. 9th International Conference on Intelligent Computing and Information Systems (ICICIS’19)

[2] AS. Keceli, A.kaya, “Violent activity detection with transfer learning method,” Electronics Letters, vol. 53, no. 15, pp. 1047– 1048, June 2017.

[3] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6, Aug 2017.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields” Apr. 2017

[5] Alexandros Stergiou and Ronald Poppe, “Spatio-Temporal FAST 3D Convolutions for Human Action Recognition”, 18th IEEE, Dec. 2019

[6] Lili Meng, Bo Zhao, Bo Chang, Gao Huang, Frederick Tung, Leonid Sigal, “Where and When to Look? Spatio-Temporal Attention For Action Recognition in Videos”, Under review as a conference paper at ICLR 2019

[7] Lei Shi, YifanZhang, Jian Cheng and Hanqing Lu, “Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition”, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12026-12035