# 인공지능을 활용한 스트리밍 서비스/SNS 내에서의 폭력 감지 시스템

김선민, 이석원, 임승수, 최상일
강릉원주대학교 컴퓨터공학과
ksmabcd135@gwnu.ac.kr actto1014@gwnu.ac.kr eungsu@gwnu.ac.kr schoi@gwnu.ac.kr

# Violence Detection System in Streaming Service and SNS Using Artificial Intelligence Technologies

Seon-Min Kim, Seok-Won Lee, Seung-Su Lim, and Sangil Choi
Department of Computer Science & Engineering, Gangneung-Wonju National University

**요 약**

인터넷 및 IT 기술의 발전과 더불어 미디어산업에도 큰 변화가 일어나고 있다. TV 를 대신하여 스트리밍 서비스를 이용하는 사람들이 늘고 있으며 SNS 를 활용하여 서로의 경험을 간접적으로 공유하는 형태의 새로운 문화 컨텐츠가 자리잡아가고 있다. 하지만 이러한 컨텐츠를 소비하는 주요 계층 중에는 초중고 학생들도 포함되어 있다. 인터넷 혹은 SNS 에서 소비되는 컨텐츠들을 관리 감독하는 컨트롤 타워가 부족하거나 전무하기 때문에 폭력, 음주, 흡연 등 사회적으로 악영향을 줄 수 있는 영상 또는 사진이 무분별하게 생산되어 청소년들에 의해 소비되고 있으며 더 나아가 이것이 사회적 문제로까지 대두되고 있다. 이러한 문제를 해결하기 위해 인공지능 기술을 활용한 여러 다양한 감시 시스템 개발을 위한 연구가 한창이다. 본 연구에서는 SNS 및 스트리밍 서비스에서 제공되는 영상 및 사진을 Pose Estimation 및 표정 인식 기술을 활용하여 폭력을 자동적으로 감지할 수 있는 폭력 감지 시스템을 개발하는데 그 목적이 있다.

## 1. 서론

최근 인터넷의 발전과 4 차산업혁명의 도래로 미디어 시장 또한 큰 변화가 일어나고 있다. 대표적으로 스포츠, 음식 등 다양한 컨텐츠를 기반으로 실시간으로 소통하거나 영상을 편집하여 특정 플랫폼에 업로드 하는 1 인 미디어의 수요가 급증하고 있으며 자신 및 타인의 근황을 공유하는 SNS 서비스가 인기를 끌고 있다. 이와 같은 공유 컨텐츠가 확산되면서 검증되지 않거나 걸러지지 않은 정보들이 무분별하게 소비되고 있는 것이 현실이다. 이러한 정보의 유통은 예기치 않은 문제들을 양산할 가능성이 매우 높다.


[그림 1] 방송 편성 과정[1]

[그림 1]은 공영방송 편성 과정을 보여준다. 공영방송의 경우 한정된 컨텐츠를 방송심의위원회의 검증과정을 거친 후 송출하기 때문에 사전에 위험 요소들을 차단할 수 있지만 스트리밍 서비스 및 SNS 의 경우 이러한 사전 작업없이 무분별하게 정보를 공유하는 경우가 대부분이기 때문에 인터넷 방송이나 SNS 에서 음주, 흡연, 폭력 등 자극적이며 사회적 문제를 발생시킬 수 있는 장면들이 아무 제재없이 송출되는 상황이 발생한다.

국내 스트리밍 서비스 시장에서 높은 점유율을 자랑하고 있는 아프리카 TV 의 경우 시스템 관리자가 직접 컨텐츠 내에서의 폭력성 여부를 판별하고 있다. 하루 동안 제작되는 대량의 컨텐츠를 관리자가 일일이 감시하여 폭력성 여부를 판단하는 것은 불가능에 가깝다.

[표 1] 플랫폼 별 분당 이용량[2]

| | 플랫폼 | | |
|---|---|---|---|
| | YouTube | Instagram | FaceBook |
| 분당 이용량 | 277 만 영상 시청 | 8611 사진 업로드 | 300 시간 영상 업로드 |

전세계적으로 수 많은 사용자들이 이용하고 있는 SNS 및 영상 제공 플랫폼인 YouTube, Instagram, Face

[1] 방송편성의 이론과 실제, 한국방송통신대학교, 2015 년 발행

[2] IDC(International Data Corporation)분석 자료

Book 의 경우 부적절한 사진 및 영상을 부분적으로나마 인공지능 기술을 이용하여 판별하고 있으나 많은 사용자들이 그 정확성에 의문을 제기하고 있다. [표 1]에서와 같이 엄청난 양의 데이터가 매일 업로드 되고 있기 때문에 이러한 방대한 양의 컨텐츠를 관리자가 직접 관리하고 확인하는 것은 불가능에 가깝다.

본 연구는 이러한 문제점을 해결하기 위하여 사람의 골격 정보를 파악할 수 있는 Pose Estimation 기술과 얼굴의 특징을 읽어 감정을 예측하는 Emotion Detection 기술을 접목하여 폭력적인 장면을 검출하는 새로운 방향을 제시하고자 한다.

## 2. 관련 연구

국내에서 상용화를 위해 시험을 마친 폭력 감지 시스템의 경우 승강기의 내의 폭력 감지 시스템, 교내 폭력 상황 감지 및 알림 시스템이 있다. 더불어 폭력 감지에 대한 연구 또한 많이 이루어지고 있다. 그렇지만 '제약된 공간에서 객체가 적게 출현해야한다'와 같이 특정 조건을 만족 시켜야 정확성이 보장되는 문제를 안고 있다.

승강기 CCTV 에서의 폭력 감지 연구[1]에서는 폭력 행위가 발생할 시에는 객체의 형태가 심하게 변한다는 점을 이용하여 화면상의 객체의 크기 및 변화 횟수가 급격히 증가하면 이를 폭력행위로 간주한다. 객체의 움직임이 한정되고 밀폐된 공간이라는 조건을 충족시키는 승강기에선 신뢰성이 높은 결과를 얻을 수 있다. 하지만 객체의 움직임이 많은 개방되어 있는 공간에서는 정확도가 떨어질 가능성을 내포한다.

두 번째 연구는 드론을 활용하여 시위 중에 발생하는 폭력을 인식하는 시스템에 관한 것이다[2]. 이 연구에서는 드론으로 시위 영상을 촬영한 뒤 서버로 데이터를 전송하여 Yolo(You Only Look once)를 사용하여 위험한 객체(총, 파이프 등)를 검출하고 폭력과 관련된 움직임을 학습시켜 폭력적인 상황을 검출한다. 이 연구에서는 위험한 객체를 검출하는 것에는 높은 정확도를 얻었지만 폭력성을 내포하는 움직임을 검출하는 데는 정확도가 떨어졌다. 정확도가 떨어진 이유는 3.1 절의에서 자세히 설명한다.

관련 연구에서 보는 바와 같이 특정 조건을 만족해야만 정확도가 높아지거나 특정 위험 객체를 검출하는 것에만 높은 정확도를 나타내는 시스템은 이와 같은 특수 상황이나 조건을 만족시킬 수 없는 스트리밍 서비스 및 SNS 에서의 폭력 감지에는 적용하기 어렵다. 따라서 스트리밍 서비스 및 SNS 를 위한 새로운 폭력 감지 시스템이 요구된다.

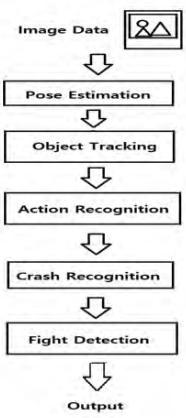
## 3. Pose Estimation

### 3.1 Pose Estimation 의 사용 이유

[그림 2]는 Object Detection 의 결과를 보여준다. 폭력 장면이 담긴 영상에서 물리적 충돌 상황을 폭력으로 정상적으로 인식하지만(왼쪽 사진) 동시에 악수와 같은 단순 신체 접촉(오른쪽 사진)도 폭력으로 잘못 탐지하는 경우가 발생하므로 객체의 관절 좌표를 이용하는 Pose Estimation 기술을 활용하여 이 문제를 해결하였다.

[그림 2] Object Detection 기반 폭력 탐지 결과

### 3.2 폭력 탐지의 과정




[그림 3] 전체적인 Fight Detection 과정

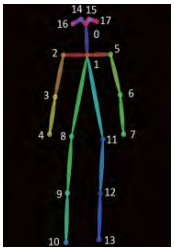
[그림 3]은 이미지 데이터에서 폭력을 감지하기 위한 전체 과정을 나타낸다. 첫째, 이미지에 있는 사람의 관절 좌표를 구하기 위해 OpenPose 를 이용하여 Pose Estimation 을 실행한다. 둘째, 이미지에 있는 사람을 실시간으로 추적하기 위하여 SORT(Simple, Online, Realtime, Tracking Algorithm) 를 이용한 Object Tracking 을 실행한다. 셋째, 실시간으로 추적되고 있는 사람의 이미지가 어떤 행위를 하고 있는지 Action Recognition 을 실행하여 4 가지 동작(서있기, 걷기, 발차기, 주먹 지르기)으로 구분한다. 넷째, Crash Recognition 을 실행 하여 이전 단계인 Action Recognition 에서 발차기 또는 주먹 지르기가 감지 된다면 Object Tracking 을 이용하여 얻은 Object 간 접촉이 있었는지 감지한다. 마지막으로 Fight Detection 단계에서 객체(사람)가 발차기 또는 주먹을 지르면서 다른 객체와도 접촉해 있다면 폭력 상황으로 판단한다. Pose Estimation, Object Tracking, Crash Recognition 단계는 OpenPose 와 SORT 라이브러리를 사용하였다. 따라서 Fight Detection 을 위한 상세 과정에 대한 내용은 Action Recognition 만을 다룬다.

### 3.3 Action Recognition 을 위한 Dataset 수집

Action Recognition 을 위해 사람의 행동을 발차기, 주먹 지르기, 서있기, 걷기 총 4 가지로 구분했다. 주먹 지르기 데이터는 MHAD(Berkeley Multimodal Human Action Database) dataset 을 이용하였다. 이 비디오 데이

터는 4 개의 각도에서 촬영된 5 개의 반복에 대해 주먹 지르기 동작을 하는 12 명을 대상으로 구성된다. 나머지 3 개의 동작은 CMU Panoptic Dataset 을 이용했다. 이 비디오 데이터는 31 가지 각도에서 촬영된 3 가지 동작을 수행하는 13 명을 대상으로 구성된다.

### 3.4 관절 좌표를 이용한 학습 데이터 변환



[그림 4] OpenPose 를 이용한 관절 좌표 정보

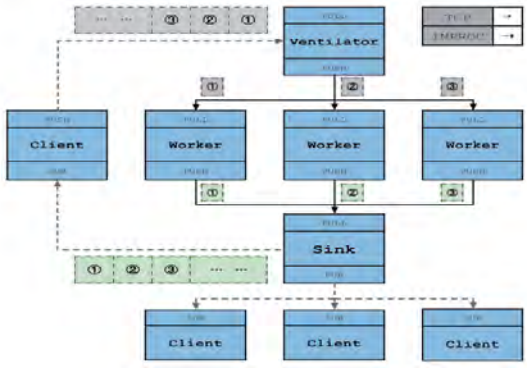
[표 2] 특징 벡터

| Index | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Angle | 2-3 | 3-4 | 5-6 | 6-7 | 8-9 | 9-10 | 11-12 | 12-13 |
| ΔAngle | 2-3 | 3-4 | 5-6 | 6-7 | 8-9 | 9-10 | 11-12 | 12-13 |
| ΔPoint | 3 | 4 | 6 | 7 | 9 | 10 | 12 | 13 |

[그림 4]는 OpenPose 를 이용한 총 18 개의 관절 좌표를 나타낸다. 이중에서 2~13 번의 관절 좌표 만을 학습 데이터로 사용했다. [표 2]는 특징 벡터를 나타낸다. 각 관절 좌표를 이용하여 관절의 각도, 관절 좌표의 변화, 관절 각도의 크기 변화, 현재 프레임에서의 관절 각도 등 총 3 가지 특징을 벡터로 변환한다. 이전 프레임의 데이터와 비교하여 각 변화의 크기를 알 수 있다.

### 3.5 딥러닝 서버

실시간에 가까운 딥러닝 연산 처리를 수행하기 위해 GCP(Google Cloud Platform)의 GPU(Nvida V100)를 대여하여 서버를 구축하였다. 서버와 클라이언트 간에는 ZMP 메시지 라이브러리를 사용하여 TCP 통신을 하도록 구현하였다(모든 메시지는 JSON 형식). 서버 내부에서는 GPU 자원을 최대한 활용하기 위해 [그림 5]와 같은 병렬 파이프 라인 구조를 적용하였다. 각 프로세스 간 통신 또한 ZMQ 메시지 라이브러리를 사용하여 구현하였으며 각 프로세스별 세부적인 기능은 [표 3]과 같다.




[그림 5] 이미지 처리 서버 구조

[표 3] 이미지 서버 구조

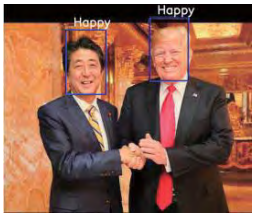
| | |
|---|---|
| **Client** | 처리할 이미지를 Server 측 Ventilator 에게 전송 |
| **Ventilator** | Client 가 전송한 이미지를 수신하여 Worker 에게 순차적으로 분배 |
| **Worker** | Ventilator 로부터 수신한 이미지 처리 작업을 수행 후 Sink 에게 전달<br>※Object Detection 기반 폭력 탐지 : YOLO 모델의 CNN 연산을 수행<br>※Pose Estimation 기반 폭력 탐지 : OpenPose 모델의 CNN 연산을 수행 |
| **Sink** | Worker 로부터 수신한 이미지 처리 작업 결과에 추가 처리 작업을 수행하고 순차적으로 Client 에게 전달<br>※Object Detection 기반 폭력 탐지 : YOLO 탐지 결과를 바탕으로 이미지에 Bounding Box 를 그림<br>※Pose Estimation 기반 폭력탐지 : OpenPose 탐지 결과를 바탕으로 SORT Object,Tracking, RNN Action Recognition,충돌 탐지 작업을 수행<br>※CNN 연산은 Darknet,RNN 연산은 Tensorflow 를 사용 |

## 4. Emotion Detection

### 4.1 표정 인식 방법

표정을 분석한 후 사람의 감정을 인식하여 분류하는 방법은 크게 Convolutional Neural Networks (CNN)을 이용한 방법과 얼굴의 특징점을 이용하는 방법이다. 본 논문에서는 위의 두가지 방법을 모두 사용하여 결과를 도출하였다.

### 4.2 CNN 을 이용한 감정 분류




[그림 6] CNN 을 이용한 표정 분류 결과

[그림 6]은 CNN 을 이용한 감정 분류[3] 결과를 보여준다. 감정 분류의 과정은 다음과 같다. 첫째, Haar Cascade 를 이용하여 이미지에서 얼굴의 위치를 찾는다. 둘째, 확인된 얼굴을 48×48 사이즈로 변환하여 CNN 에 입력한다. 셋째, CNN 실행 결과를 7 가지 감정(Angry, Disgusted, Happy, Neutral, sad, surprised, fearful)에 대한 Sotfmax 점수를 출력한다. 마지막으로 가장 높은 점수를 받은 감정을 화면으로 출력한다. 그러나 CNN 을 이용한 감정 분류는 첫번째와 세번째 단계에서 제한적인 학습 데이터로 인해 카메라를 정면으로 보고 찍은 사진만 분류가 된다는 한계가 있다. 포착된 얼굴이 조금이라도 틀어진 모습이 있으면 이미지 상에서 얼굴을 찾지 못하고 결과적으로 표정을 감정별로 분류할 수 없다.

### 4.3 얼굴의 특징점을 이용한 감정 인식

OpenPose 의 Face Tracking 은 얼굴의 특징점을 찾는

프로그램이다. [그림 7]은 Face Tracking 의 예시이며 정면의 얼굴뿐만 아니라 카메라를 기준으로 각도가 틀어져 있는 얼굴에 대해서도 얼굴의 특징을 포착하는 것을 볼 수 있다. 또한 얼굴의 특징점 좌표와 KNN(K-Nearest Neighbors Algorithm)을 이용하여 사용자 감정 인식을 성공적으로 수행한 사례도 있다[4].

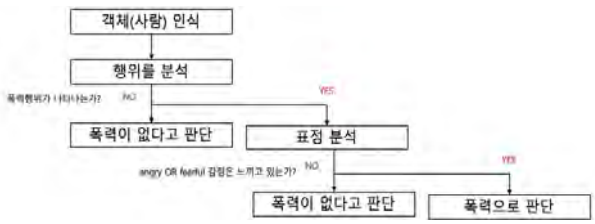
[그림 7] OpenPose 의 Face Tracking 예시

[그림 8] 싸움 이미지 판독 결과

특징점이란 얼굴의 윤곽, 눈, 눈썹, 코, 입을 인식하여 점으로 나타낸 것을 말하며 이 특징점들이 모두 연결되어 [그림 7]과 같이 윤곽선이 잡힐 경우 얼굴이 정확하게 인식되었다고 한다. [그림 7]과 [그림 8]은 영상을 OpenPose 의 Face Tracking 을 이용하여 실행한 결과이다. [그림 7]의 경우 감정 인식을 위해 사용되는 특징점들이 정확하게 인식된 반면, [그림 8]의 경우 폭력을 행하는 사람은 얼굴 윤곽선, 코, 입, 눈에 대한 특징점 중 윤곽선 및 눈썹에 대한 특징점이 누락되었고 폭력을 당하는 사람의 특징점은 하나도 검출되지 못하였는데 이것은 [그림 8]의 영상이 [그림 7]의 영상에 비해 상대적으로 빠른 움직임과 데이터 학습 부족으로 인해 특징점을 검출하지 못하여 생긴 현상이며 이러한 특징점 누락으로 인해 감정 인식에 실패하여 결과적으로 폭력 상황을 검출하지 못했다. 스트리밍 서비스 특성상 [그림 8]과 같이 움직임이 많은 상황이 대부분이기 때문에 Face Tracking 만을 이용하여 폭력성을 인지하는 것은 불가능하다고 판단된다.

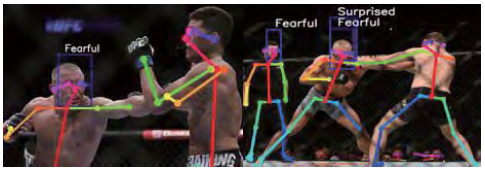
## 5. 해결방안

[그림 9]는 해결방안으로 제시한 폭력을 검출하는 알고리즘을 나타낸 것이다. 먼저 행위를 분석하여 폭력적인 모습이 나타나는지를 확인한 후 폭력적인 장면이 인식되면 그 다음 단계에서 표정을 인식하여 사람의 감정을 판독한다. 판독 결과에서 Fearful 또는 Angry 라는 감정을 인식하면 결론적으로 폭력적인 상황이 발생했다고 결론을 내린다.




[그림 9] 폭력을 검출하는 알고리즘

[그림 10]은 [그림 9]에 나타나있는 감정과 행위를 동시에 감지하는 알고리즘을 통해 감지해낸 결과이다. 행위 분석 또는 표정을 통한 감정 분석 중 한 가지의 기술을 이용하여 폭력을 검출했을 경우 단순한 접촉을 폭력으로 판단하거나 얼굴의 각도가 빠르게 변화하여 감정 검출에 실패하여 폭력을 검출하지 못했던 영상, 사진에 대하여 Fearful, Surprise Fearful 이란 감정을 추출함으로써 폭력적인 상황이 발생하고 있다고 판단한 모습을 보여준다.




[그림 10] 감정과 행위 감지를 함께 실행한 예시

## 6. 결론

폭력적인 상황의 특성상 폭력을 행하고 있는 사람과 폭력을 당하고 있는 사람 모두 화면에 얼굴을 정면으로 향하게 하는 것은 쉽지 않다. 발차기와 주먹지르기 행위를 감지하는 동시에 사람의 감정을 감지하여 폭력을 판단 하려면 사람의 정면 얼굴은 물론이고 부분 얼굴 표정을 통한 감정을 파악하는 과정이 필수적이다. 이를 위해 지금까지 2 가지 방법을 사용했다. 그러나 학습데이터의 한계, 불안정한 얼굴 특징점 파악이라는 두가지 문제가 있다. CNN 을 활용하여 개선하는 방법으로 정면 얼굴의 감정 인식은 기존 시스템을 이용하되 부분 얼굴(profile face)은 LBP Cascade Classifier 을 이용하여 이미지에서 부분얼굴인식을 거친 후 감정 분류를 진행할 수 있다. 그러나 감정 분류를 위해 CNN 을 학습시키려면 각 감정에 대한 부분 얼굴 DataSet 을 충분히 확보해야한다. 하지만 이러한 문제점이 있음에도 불구하고 하나의 기술만을 사용하였을 경우에는 실제로 폭력적인 상황이 발생했지만 폭력이라고 인식하지 못했던 문제점이 2 가지 이상의 기술을 접목한 후 정상적으로 폭력으로 인식하는 것을 확인 할 수 있었다. 이러한 기술을 SNS 및 스트리밍 서비스와 접목시킨다면 폭력적인 장면이 송출되거나 사진이 업로드 되는 상황을 빠르게 막을 수 있을 뿐만 아니라 거대한 용량의 데이터를 인간이 처리하는 것보다 빠르고 정확하게 처리할 수 있을 것으로 예상한다.

## 참고문헌

[1] 심영빈, 박화진, “CCTV 에서 폭력 행위 감지 시스템 연구”, 디지털 컨텐츠 학회, 제 16 권, 제 1 호, pp 26-30, 2015.

[2] Yeon-Su Lee, Hyun-Chul Kim, “Deep Learning-based Violent Protest Detection System”, 한국컴퓨터 정보학회논문지, 제 2 권, 제 3 호, pp 87-93, 2019.

[3] Enrique Correa, Arnoud Jonker, Michael Ozo, Rob Stolk, “Emotion Recognition using Deep Convolutional Neural Networks”, Tech Report IN4015, TU Delft, 2016.

[4] 이용환, 김홍준, “얼굴 특징점 추적을 통한 사용자 감성 인식”, 반도체디스플레이기술학회지 제 18 권, 제 1 호, pp 98-100, 2019.