JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 4, pp. 1069-1076, Apr. 2025

Check for updates

이상 행동탐지기반의 자동 선별적 비식별화 연구

김 대 진[1] · 전 윤 걸[2] · 김 준 화[3*]
[1]동국대학교 영상문화콘텐츠연구원 조교수
[2]성균관대학교 체력과학연구소 연구원
[3]건양대학교 인공지능학과 조교수

Automatic Selective De-Identification Based on Abnormal Behavior Detection

Dae-Jin Kim[1] · Youn-Girl Jeon[2] · Jun-Hwa Kim[3*]

[1]Assistant Professor, Research Institute for Image & Cultural Contents, Dongguk University, Seoul 04626, Korea
[2]Researcher, Research institute for Physical Fitness and Sport Science, Sungkyunkwan University, Suwon 16419, Korea
[3]Assistant Professor, Department of Artificial Intelligence, Konyang University, Daejeon 35365, Korea

[요 약]

지금까지는 자동으로 비식별화하는 경우에 얼굴인식이나, 객체탐지를 이용하여 프레임 단위의 비식별화를 하고 있다. 그러나, 사람들이 거부감을 가질 수 있는 상황(혐오, 폭력 장면 등)에 대해서도 동작 단위의 비식별화가 요구되고 있기 때문에 이러한 경우 동영상 편집기를 이용하여 수작업으로 비식별화를 진행하고 있다. 본 논문에서는 동작 단위의 비식별화를 위해서. 공간영역의 객체탐지(YOLOv5)와 동작 인식(MViT) 기술을 이용하고, 시간 영역에서 트랜스포머 기반의 이상 탐지를 통해서 비식별화 대상이 되는 프레임의 시작과 끝 시점을 분석하여, 해당 프레임 섹션에 비식별화를 진행하였다. 본 실험에서는 다수의 동영상을 비식별화 하였고, 이상행동 5종(쓰러짐, 폭행, 주저앉음, 사고, 기물파손)에 대해서 평균 72.8%의 정확도를 가진다.

[Abstract]

Until now, in automatic de-identification, face recognition or object detection has been used to de-identify a frame. However, de-identification of motion units is required in anomal situations (disgust, violence, etc.), necessitating video editing. In this paper, we propose a method for de-identification of motion units, using you only live once (YOLOv5) for object detection and multiscale vision transformer (MViT) technology for motion recognition in the spatial domain, accompanied by transformer-based anomaly detection in the temporal domain, to analyze the start and end of the corresponding frame sections to be de-identified. In this experiment, we de-identified a large number of videos, achieving an average accuracy of 72.8% for five types of anomalous behavior (falling down, assault, sitting down, accident, and vandalism).

색인어 : 선별적 비식별화, 이상 탐지, 개인정보, 동작 인식, 객체 인식
Keyword : Selective De-Identification, Anomaly Detection, Personal Information, Action Recognition, Object Detection

http://dx.doi.org/10.9728/dcs.2025.26.4.1069

Received 24 February 2025; **Revised** 17 March 2025
Accepted 25 March 2025

***Corresponding Author; Jun-Hwa Kim**

Tel: +82-42-600-8508
E-mail: junhwakim@konyang.ac.kr

Ⅰ. 서 론

최근 출입통제가 필요한 곳뿐만 아니라, 거리, 지하철역과 같은 공공장소에서도 CCTV(Closed Circuit Television)는 어디서든 볼 수 있는 상황이 되었다. 이러한 상황은 개인의 안전 및 보호를 위해서 필수적인 요소지만, 비디오 감시 때문에 저장된 영상들에 의해서 의도하지 않은 개인정보가 수집될 수 있으므로 개인정보 보호의 문제를 발생시킬 수 있다[1]. 이러한 이유로 저장된 영상이 외부로 배포가 되는 경우에 비식별화가 필수적으로 동반되어야 하며, 이를 미흡하게 하여 특정 개인을 식별하게 되는 등의 안전 의무를 위반한 경우 5년 이하의 징역 또는 5,000만원 이하의 벌금 및 전체 매출액의 3%에 해당하는 과징금을 부과하는 등 개인 비식별화에 의한 사생활 보호 기술의 중요성이 매우 높아지고 있다[2]. 개인정보의 보호를 위해 비식별화하는 경우도 있지만, 이 외에도 사람들에게 거부감을 줄 수 있는 영상에 대해서도 비식별화가 이루어져야 한다. 예를 들어 뉴스에서 CCTV로 찍힌 영상 중 폭력성이 짙거나, 혐오감을 불러일으키는 영상에 대해서도 비식별화를 진행하여 방송에 내보낸다. 이는 시청자들이 영상 내 장면이 비윤리적이고, 범죄 조장을 할 수 있으며, 그 심각성으로 인해 폭력과 혐오문제를 일으킬 수 있으므로 비식별화가 반드시 필요하다. 이러한 문제점을 줄이기 위해서 본 논문에서는 컴퓨터 비전기술을 기반으로 하여 프레임 내의 지정 객체뿐만 아니라 사람들이 거부감을 가질 수 있는 상황에 대해서도 동작 단위의 비식별화를 할 수 있는 이상 탐지기반의 자동 선별적 비식별화 기법에 관한 연구를 한다. 본 논문의 구성은 2장에서 비식별화를 위한 기존 사례 및 문제점을 제시하고 3장에서는 이상 탐지기반의 비식별화 알고리즘을 알아보고, 영상 내 비식별화시 이상 탐지 알고리즘을 어떻게 연동할 수 있는지를 연구하며, 4장에서는 제안한 알고리즘을 기반으로 구현해 보고, 주어진 환경에서의 성능 비교를 하였으며, 5장에서는 결론을 지었다.

Ⅱ. 기존 사례 연구

2-1 얼굴탐지 기반의 비식별화

프레임 기반의 대표적인 비식별화 방식인 얼굴탐지 기반의 비식별화 프로세스는 그림 1과 같다

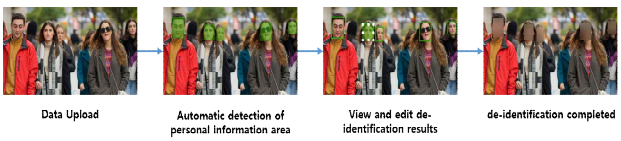

그림 1. 얼굴탐지 기반의 비식별화[1]
Fig. 1. De-identification based face detection[1]

1) 데이터 업로드

대용량의 사진 및 동영상을 FTP(File Transfer Protocol), HTTP(Hypertext Transfer Protocol) 등 전송 프로토콜을 이용해서 비식별화 처리 서버로 업로드한다.

2) 개인정보 영역 자동 탐지

영상을 통해서 개인정보를 확인할 수 있는 얼굴 영역을 AI(Artificial Intelligence) 알고리즘을 이용하여 자동으로 탐지한다. 최근에 AI 기반의 고속 얼굴탐지 기술들이 많이 발달하였고, 대표적인 알고리즘으로 MTCNN[3], ArcFace[4], RetinaFace[5], CenterFace[6] 등을 사용할 수 있다. 알고리즘 학습 시 LFW(Labeled Faces in the Wild)[7], WIDER FACE[8]등 비즈니스 모델에 맞추어서 적합한 데이터셋을 결정하고 이 학습된 데이터를 통해서 얼굴탐지에 사용한다.

3) 비식별 처리 결과 확인 및 편집

비식별 처리된 결과물을 확인하고 수정이 편집 도구 등을 이용하여 해당 영역을 수정한다.

4) 비식별 처리 완료

비식별화가 완료 된 후에 마스킹이 된 형태로 데이터를 엔코딩(Encoding)하여 영상을 재구성한다.

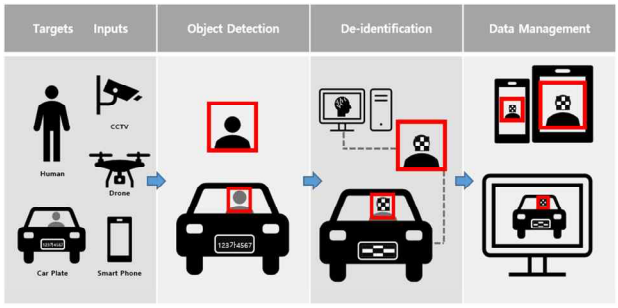

그림 2. 객체탐지 기반의 비식별화[1]
Fig. 2. De-identification based object detection[1]

2-2 객체 탐지 기반의 비식별화

일반적인 객체탐지 기반의 비식별화 프로세스는 그림2와 같다. 특히 얼굴의 경우 출입통제 시 사용이 될 수 있는 고유의 개인정보이고, 차량의 경우 차량번호에 의해서 소유주 정보를 알 수 있는 개인화된 정보 영역이다. 얼굴탐지의 경우 2-1과 같이 비식별화를 할 수 있고, 차량번호의 경우는 객체탐지를 이용하여 번호판(Plate)을 찾아 비식별화를 할 수 있다. 이때 객체탐지를 위해서 YOLOv5[9], SSD[10], Faster R-CNN[11] 등과 같은 인공지능 알고리즘을 사용하며, 비즈니스 모델 및 실시간성 여부에 따라서 알고리즘이 달리 사용할 수 있다. 특히 실시간성이 필수인 경우 YOLO, SSD와 같

은 One Stage Model을 사용하며, 속도는 느리지만, 정확성이 더 필요한 경우 Faster R-CNN과 같은 Two Stage Model을 사용하여 비식별화 진행할 수도 있다.

2-3 문제점 제시

기존의 모든 비식별화 솔루션은 지정 객체 또는 사람 얼굴에 대해서만 비식별화가 이루어지고 있다. 이러한 경우는 개인의 권리 보호를 위해서 비식별화가 이루어진 것이다. 그러나, 비식별화의 대상은 프레임 내의 지정 객체뿐만 아니라 사람들이 거부감을 가질 수 있는 상황(혐오, 폭력 장면 등)에 대해서도 동작 단위의 비식별화가 필요하다. 따라서, 이러한 문제를 해결하기 위해서 영상에 대한 이상 탐지를 진행하고, 탐지된 시간 영역에 대해서 자동 선별적 비식별화를 진행하여 이러한 문제를 해결할 수 있다.

Ⅲ. 이상 탐지기반의 비식별화 연구

3-1 Spatial Domain 비식별화 기법연구

1) 객체탐지(Object Detection)

객체탐지의 경우 실시간성 또는 정확성 중 어느 부분이 더 중요한가에 따라 선택하는 인공지능 모델이 다르다. 실시간성이 중요한 비즈니스 모델에서는 One Stage Detector를 사용하고, 정확도를 위해서는 Two Stage Detector를 많이 사용한다.

본 논문에서는 실시간성을 더 중요하게 여기기 때문에 대표적인 모델인 YOLO(You Only Look Once)v5를 사용한다. 이 모델의 특징을 보면 Neural Network Single Forward만으로 다수의 객체탐지가 가능하고, 빠르게 탐지할 수 있으며, 이미지 전체 Context로 탐지할 수 있으므로, 상대적으로 False Positive가 적게 발생한다.

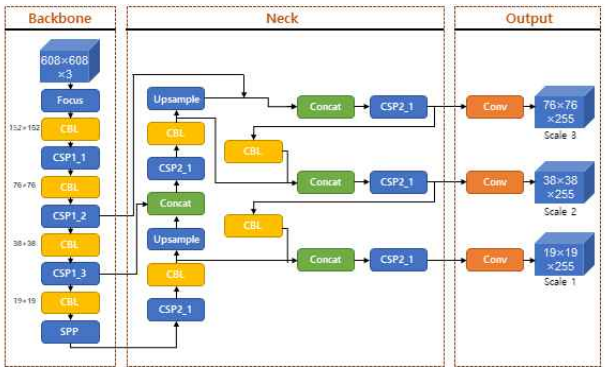

그림 3. YOLOv5 네트워크 모델
Fig. 3. YOLOv5 network model

YOLOv5의 모델 네트워크를 보면 그림 3과 같다. YOLO 5 알고리즘은 크게 3개의 부분(Backbone, Neck, Output)으로 구성된다. Backbone에서 입력 영상의 Feature 정보를 추출하고, Neck은 추출된 Feature 정보기반으로 특징 맵의 3개의 축척을 생성한다. 마지막으로 Output 부분에서는 생성된 기능 맵을 통해서 객체탐지를 하도록 구성된다.

Backbone은 Feature Map을 추출하는 Convolution Network로 다중 Convolution과 Pooling을 통해서 4개의 Map Layer를 구성한다. 이때 152×152, 76×76, 38×38, 19×19픽셀 Layer를 생성한다.

Neck에서 더 많은 Context를 얻고 Loss 값을 줄이기 위해서 Backbone으로부터 받은 Map Layer는 FPN(Feature Pyramid Network)과 PAN(Path Aggregation Network)의 피라미드 구조를 사용할 수 있다. FPN은 상위 Feature Map에서 하위 Feature Map까지 모든 의미 있는 Feature 정보를 가지고, PAN은 하위 Feature Map으로부터 강인한 Localization Feature 정보를 상위 Feature Map에 전달하여 융합 Feature 정보를 구성한다. Feature의 정보는 76×76×255, 38×38×255, 19×19×255차원의 융합 Layer를 가진다. 다차원이 융합 Layer에서 탐지 과정에 의해서 크고 작은 모든 객체에 대해서도 SOTA(State of the Art)의 결과가 나온다.

YOLOv5는 s, m, l, x, 4가지 모델로 제공되며 서로 다른 감지 정확도와 성능을 제공한다. 이는 비즈니스 모델에 따라 가장 적합한 모델을 사용하면 된다. 본 논문에서는 속도에 중요성을 강조한 s 모델을 기반으로 테스트하였다.

비식별 대상을 찾기 위해서는 객체탐지가 반드시 이루어져야 한다. 특히 비식별화의 대상이 혐오, 폭력 등의 장면인 경우 대부분 사람과 직접 연관된 것이다. 따라서 이상행동으로 판단되면 해당 프레임 내에서 사람에 대한 객체탐지를 진행하고 비식별화를 진행한다.

2) 동작인식(Action Recognition)

기존의 동작 인식 방식들은 대부분 전체 비디오를 분석하거나 각 세그먼트에 대해서 단일 작업 레이블을 할당하여 동작 인식을 한다. 이때, 우리(인간)는 장면을 인식하기 위해 시각적 데이터의 인스턴스만 필요하다는 것을 추론할 수 있다. 즉, 작은 프레임 그룹 또는 영상의 단일 프레임만으로도 정확한 인식에 충분하다. 이러한 방법을 이용하면 영상 내 연속 스트림에서 얻은 프레임을 기반으로 관심 있는 동작을 감지 및 인식하는 접근 방식이 가능하다. 이렇게 프레임 기반의 동작 인식의 경우 객체탐지에서 사용하였던 YOLOv5를 이용하여 동작 인식 분석에 사용할 수 있다[12].

3-2 Time Domain 비식별화 기반연구

1) 동작 인식(Action Recognition)

기존의 영상 분석/인식의 경우 CNN(Convolution Neural

Network)을 기반으로 하는 많이 연구됐으나 최근에 NLP (Natural Language Processing) 분야에서 발전되어 Vision 분야에 적용된 트랜스포머(Transformer)가 다양하게 연구되고 있다. 이중 대표적인 Vision Transformer[13] 모델의 경우 인식률에서는 기존의 CNN 모델들보다 더 좋은 결과를 나타내지만, Self-Attention block에서 scale에 따라 연산량 및 메모리가 상당히 증가하는 단점을 가진다. 따라서 트랜스포머가 가지는 인식률의 장점과 리소스 부분의 최적화를 위해 연구한 모델이 MViT(Multiscale Vision Transformer)[14]이다. 기존의 Vision Transformer에서 MHA(Multi-Header Attention)를 이용하여 Encoder를 구성하였으나, MViT에서는 MHA를 Pooling Attention으로 대체하였다. 그림 4에서는 MViT의 핵심 영역인 Pooling Attention을 나타낸다.

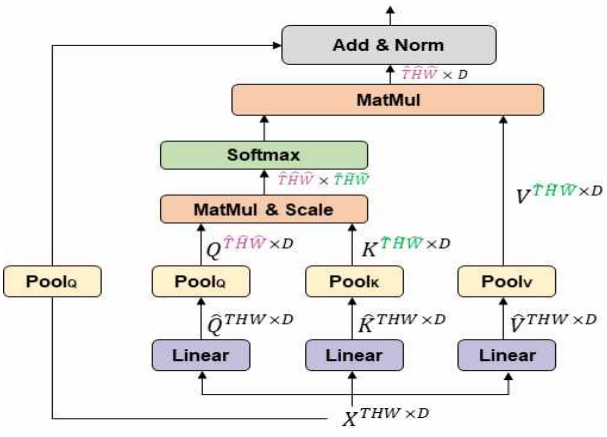

그림 4. 풀링 어텐션
Fig. 4. Pooling attention

MViT는 각 stage 별로 resolution을 줄이고 channel을 확대하는 방법으로 각 Stage에 Pooling Attention이 적용된다. input sequence인 $X \in \mathbb{R}^{L \times D}$에서 $W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$을 통한 linear projection 후에 pooling operator(P)를 통해 아래와 같이 적용할 수 있다.

$$Q = P_Q(XW_Q), K = P_K(XW_K), V = P_V(XW_V) \quad (1)$$

$P_Q(XW_Q)$을 통해 Pooling을 거친 output length $\tilde{L}$은 $Q \in \mathbb{R}^{\tilde{L} \times D}$로 줄어들며 Self Attention 연산을 수행한다.

$$Attn(Q,K,V) = Softmax(QK^T/\sqrt{D})\,V \quad (2)$$

MViT 구조를 보면 3D CNN처럼, 깊어질수록 spatial dimension은 줄이되 channel dimension은 늘려나간다. 이를 위해 Scale stage를 4단계로 나눈다. 그림 5는 MViT의 기본모델 레이어를 설명한다. data layer에서는 프레임 τ 간격으로 샘플링 한다.

stages	operators	output sizes
data layer	stride $\tau \times 1 \times 1$	$D \times T \times H \times W$
cube_1	$c_T \times c_H \times c_W$, D stride $s_T \times 4 \times 4$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
scale_2	$\begin{bmatrix} \text{MHPA}(D) \\ \text{MLP}(4D) \end{bmatrix} \times N_2$	$D \times \frac{T}{s_T} \times \frac{H}{4} \times \frac{W}{4}$
scale_3	$\begin{bmatrix} \text{MHPA}(2D) \\ \text{MLP}(8D) \end{bmatrix} \times N_3$	$2D \times \frac{T}{s_T} \times \frac{H}{8} \times \frac{W}{8}$
scale_4	$\begin{bmatrix} \text{MHPA}(4D) \\ \text{MLP}(16D) \end{bmatrix} \times N_4$	$4D \times \frac{T}{s_T} \times \frac{H}{16} \times \frac{W}{16}$
scale_5	$\begin{bmatrix} \text{MHPA}(8D) \\ \text{MLP}(32D) \end{bmatrix} \times N_5$	$8D \times \frac{T}{s_T} \times \frac{H}{32} \times \frac{W}{32}$

그림 5. MViT 기본모델 레이어[15]
Fig. 5. Multiscale Vision Transformers base model[15]

cube1은 embedding layer로 dimension을 96으로 줄였다. scale2~5는 transformer layer로 embedding 하는 패치들의 spatial resolution은 줄여나가면서 embedding dimension은 2배씩 차원 수를 높여준다. 이때 query pooling, key/value pooling, skip connection/positional encoding을 통해서 네트워크를 구성한다.

영상에 대해서 MViT를 이용하여 동작 인식을 적용하기 위해서는 다음과 같은 과정이 필요하다.

(1) 2D patch가 아닌 spacetime을 고려하여 cube token인 3D로 token을 생성함
(2) pooling operator도 spacetime을 고려한 feature map으로 처리함
(3) relative positional embedding도 spacetime을 고려한 정보로 H+ W+ T로 decompose 하여 처리함.

본 실험에서는 MViT-B 모델을 사용하였고, τ=4, cT=3, cH=7, cW=7, D=96, sT=2 의 값을 사용하였다.

3-3 이상 행동탐지 기반의 자동 선별적 비식별화

개인정보 보호, 상표 노출 등을 이유로 영상 콘텐츠 내의 개인정보에 대해 비식별화가 이루어지고 있다. 이러한 경우는 권리 보호를 위해서 비식별화가 이루어진 것이다. 또한, 그 외에도 시청하는 사람들이 거부감을 가질 수 있는 상황(혐오, 폭력 장면 등)에 대해서도 비식별화가 필요하다. 후자의 경우는 이상 탐지를 통해서 비식별화를 진행할 수 있다. 진행 과정을 보면 이상행동 선별 분석, 비식별화 시간 영역 분석, 비식별화 대상 적용으로 3단계를 거쳐서 선별적 비식별화를 진행할 수 있다. 그림 6에서는 3단계를 통해서 이상 탐지기반의 자동 선별적 비식별화 과정을 보여준다.

1) 이상 행동별 선별 분석

이상행동 탐지를 위해서 세그먼트/프레임 단위의 이상행동

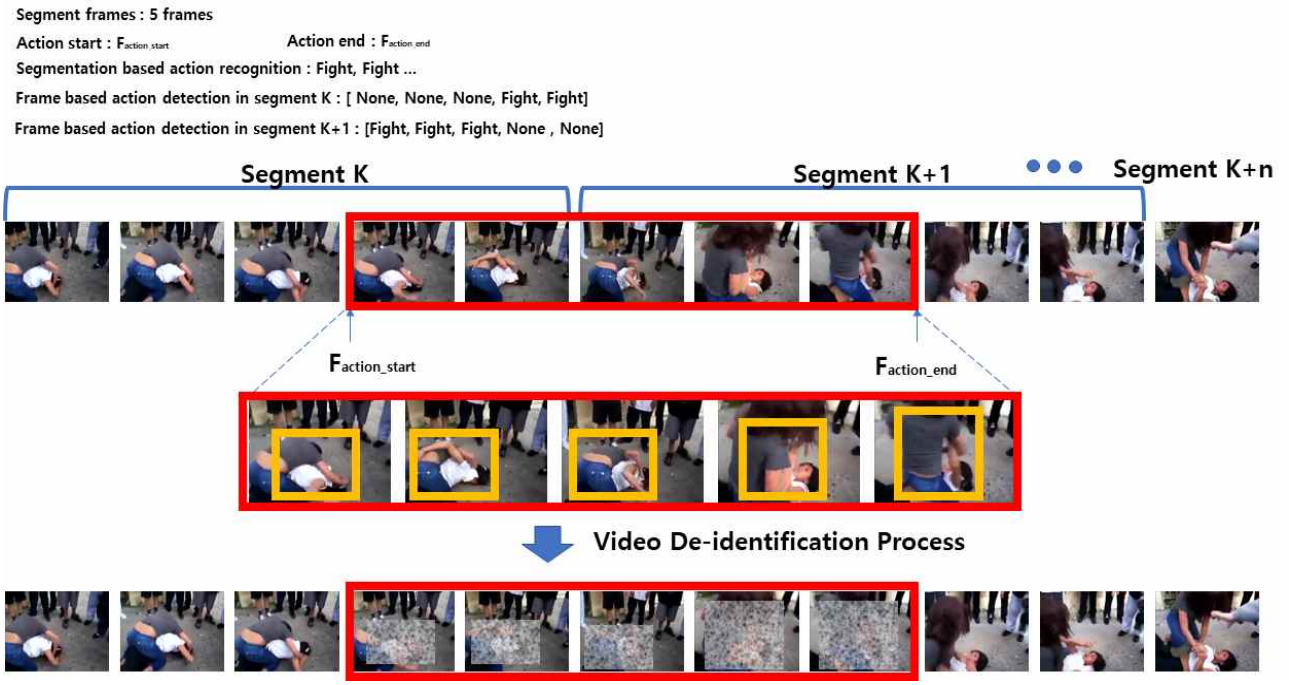

그림 6. 이상 행동탐지 기반의 자동 선별적 비식별화
Fig. 6. Anomaly detection-based automatic selective de-identification

분석을 진행한다. 우선 세그먼트 단위의 이상행동은 영상 내에서 이상행동이 일어난 위치를 먼저 선정한 후, 프레임 단위의 이상행동 분석을 통해서 세그먼트 내의 이상행동의 시작/종료 시점의 프레임들을 선정하여 해당 구간의 정밀한 비식별화를 진행할 수 있다. 이때, 세그먼트 단위의 이상행동 탐지는 MViT 모델을 사용하였고, 프레임 단위의 이상행동 탐지는 YOLOv5 모델을 사용하였다.

• **세그먼트 단위 이상행동 탐지**

MViT 행동 인지 모델을 통해서 이상행동(쓰러짐, 폭행, 주저앉음, 사고, 기물파손)과 일반행동(걷기, 달리기, 서 있기, 의자 앉기)을 검출하기 위해서 각 행동에 대해서 10초 단위로 행동을 구분하여 동영상으로 만들었고, 각 class 당 500개의 동영상으로 학습하였다. 이때 학습 인자값으로 batch=4, epoch=200, resolution=224, segment frames=5, sampling rate=3, input channel num=3을 사용하였고, MViT 모델 인자값으로 depth=24, head=1, embedding dimension=96, patch kernel=(3, 7, 7), patch stride=(2,4,4), patch padding=(1,3,3), mlp ratio=4.0를 사용하였으며, solver 인자값으로 lr=0.0001, momentum=0.9, weight decay=0.05를 사용하였으며, mode loss function으로는 soft cross entropy를 사용하고, dropout rate를 0.5로 지정하였다.

• **프레임 단위 이상행동 탐지**

YOLOv5 행동 인지 모델 학습 시에 세그먼트 단위의 이상행동 탐지와 마찬가지로 이상행동 5종과 일반행동 4종의 class로 구분할 수 있도록 하였으며, 각 class 당 3000개의 이미지를 사용하여 학습하였다. 이때 학습 인자 값으로 batch=64, epoch=100, width=416, height=416, channels=3, momentum=0.949, decay=0.0005, lr=0.001을 사용하였다.

2) 비식별화 시간 영역 분석

영상 내에서 세그먼트 단위의 이상행동 탐지결과와 세그먼트 안에서 프레임 단위의 이상행동 탐지결과가 같은지 확인하고, 탐지결과가 같은 현재 세그먼트(Segment K) 내의 첫 번째 프레임을 $F_{action.start}$로 하고, 다음 세그먼트(Segment K+1)내에서 세그먼트 단위의 이상행동 탐지결과와 세그먼트 안에서 프레임 단위의 이상행동 탐지결과가 같은 마지막 프레임을 $F_{action.end}$로 지정한다. 만약 그 다음 세그먼트(Segment K+2)에도 같은 결과이면 $F_{action.end}$는 그다음 세그먼트(Segment K+2)에 설정할 수 있으며, 이것은 N 세그먼트(Segment K+N)까지 확장할 수 있다.

3) 비식별화 대상 적용

프레임 단위 이상 행동탐지를 판단하기 위해서 YOLOv5를 사용하기도 했지만, 객체탐지를 위해서도 같은 인공지능 모델을 사용할 수 있다. 학습 이미지가 객체를 나타내느냐, 상황을 나타내느냐의 차이가 있을 뿐이고 학습데이터를 어떻게 구성하느냐에 따라서 다양한 분야에 적용할 수 있다.

비식별화를 시키는 대상이 혐오, 폭력 장면 등의 장면이므로 대부분 사람과 직접 연관된 것이다. 따라서 비식별화 시간

영역인 Frames Section 내의 객체 인식(사람)에만 적용한다. 이때 사람을 인지하도록 COCO 데이터 세트의 사람 데이터를 통해서 학습하였고 비식별 대상을 선정하였다.

4) 전체 시스템 연동

비디오 스트림 내에서 비식별화를 진행하기 위해서 비식별 대상을 연속된 스트림 내에서 찾는 것이 중요하다. 이를 위해서 세그먼트 단위(5프레임)로 이상행동 탐지를 진행하고, 이 5프레임에 대해서 각각 프레임 단위 이상 행동탐지를 진행한다. 이때 각각의 결과가 0.7의 임계값을 넘어 신뢰성을 가진다면 class로 선정한 행동으로 인지한다. 이를 통해 정해진 $F_{action.start}$와 $F_{action.end}$ 정보를 통해서 비식별화 대상의 Frames Section을 만들고 객체탐지를 통해서 사람에 대한 BBox(Bounding Box)정보를 가져오고 마스킹(Masking)을 통해서 비식별화를 진행한다. 그림 7에서는 이상 행동탐지 기반의 자동 선별 비식별화 순서도를 나타낸다.

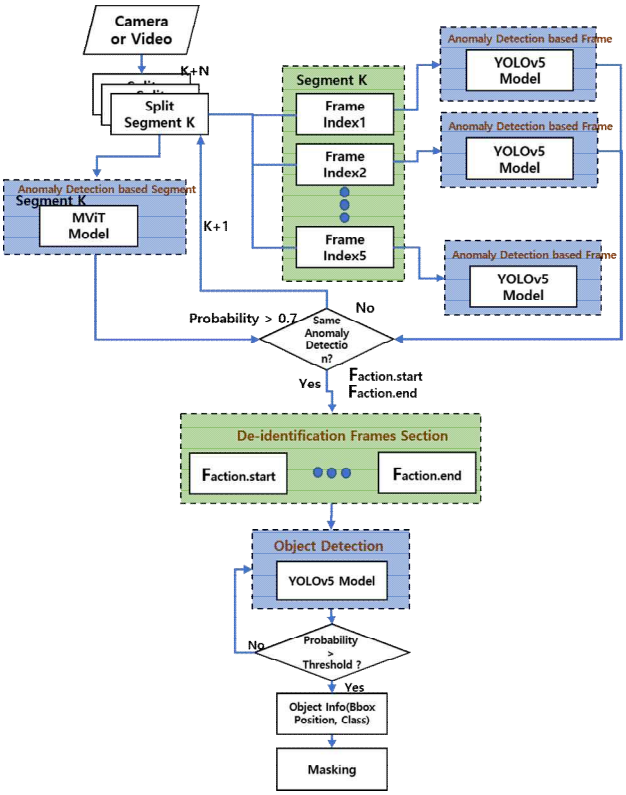

그림 7. 이상 행동탐지 기반의 자동 선별적 비식별화 순서도
Fig. 7. Anomaly detection-based automatic selective de-identification flowchart

Ⅳ. 성능측정 결과

이상 행동탐지 기반의 자동 선별적 비식별화 테스트를 위해서 30fps가 되는 3분~5분 길이의 동영상 100개에 대해서 테스트를 진행하였다. 비식별화 장비는 OS가 64bit Windows 10에 Intel Xeon Processor(Skylake, IBRS) 2.59 GHz(16 Core), 메모리는 160Gbyte를 사용하였으며, 본문에서와 같이 이상 행동탐지를 위한 인공지능 모델로 MViT, YOLOv5를 사용하였고, 객체탐지 관련 알고리즘인 YOLOv5를 적용하였으며, 데이터 학습 시 NVidia Tesla V100-PCIE-32GB Graphic Card를 사용하였다.

비식별화의 성능측정을 위해서 GT(Ground Truth)는 10명의 사람이 동영상에서 이상행동이 일어나는 구간을 각각 선정하였고, 10명의 결과 중 공통부분만 이상행동의 GT로 결정하였다.

결정된 GT를 기준으로 비식별화의 정확도를 측정하였다. 이때 비식별화 정도를 측정하는 기준 Metric은 표준화된 것이 없다. 프레임 기반 이상 탐지를 mAP, 동작 인식 기반의 이상 탐지를 Top-K Accuracy를 이용하여 측정하는 것은 비식별화 정도를 모두 포함하기에 적합하지 않기 때문에 비식별화 정도를 판단하기 위해 새로운 Metric을 사용하였다. 비식별화는 아래와 같이 이루어졌을 때 가장 정확하게 비식별화가 이루어졌다고 판단하였다.

$$F_{action.start} \leq \text{Frames Section} \leq F_{action.end} \quad (3)$$

$F_{action.start}$에서 벗어난 프레임 수를 α라 하고, $F_{action.end}$에서 벗어난 프레인 수를 β라 하면 α+β 값에 따라 정확도를 구분하였다. 여기서 100_{GT}는 GT의 일치 정확도를 나타낸다.

$$De-identification\ \ Accuracy = 100_{GT} - \frac{(\alpha+\beta)}{2} \quad (4)$$

이 기준으로 하여 100개의 동영상에 대해서 De-identification Accuracy가 이상행동 5종(쓰러짐, 폭행, 주저앉음, 사고, 기물파손)에 대해서 평균 72.8%의 정확도를 가진다. 일반행동 4종은 비식별화 대상이 아니므로 일반행동으로 인식되는 경우에는 비식별화에서 제외하였다. 대상에 대한 오인식에 따른 잘못된 비식별화 처리도, fps가 30프레임인 경우 Frames Section의 앞뒤로 1초 이내의 오차범위로 비식별화가 이루어짐을 확인할 수 있다. 이는 이상행동 경계범위에 대해서 추가적으로 비식별화가 이루어진 것이다. 비식별화의 특성상 비식별화 처리가 진행되더라도 배포에는 큰 영향을 주지 않는다. 사람이 영상을 보고 경계범위를 임의로 판단하여 비식별화 처리를 진행하더라도, 추가적인 비식별화는 이루어질 수 있다. 이러한 이유로 이상행동이 일어나지 않은 모든 프레임까지 비식별화를 진행할 경우 많은 리소스의 낭비를 가져올 수 있다. 따라서, 수작업으로 진행되는 비식별화작업을 자동비식별화 기술을 이용하여 상당량의 업무를 줄일 수 있다는 측면에서 의미가 있다.

V. 결 론

비식별화의 대상이 되는 것은 개인정보뿐만 아니라 폭력성, 혐오감을 불러일으키는 영상에 대해서도 비식별화가 이루어져야 한다. 본 논문에서는 이러한 상황에 대해서 동작 단위의 비식별화를 할 수 있는 자동 선별적 비식별화 기법에 관한 연구를 진행하였다. 시청에 비적합한 영상이 비식별화되지 않고 배포 시 비윤리적이고, 범죄 조장을 할 수 있으며, 그 심각성으로 폭력과 혐오문제를 일으킬 수 있기 때문에 자동으로 선별적 보호를 할 수 있는 연구로서 의미가 있을 수 있다. 추후에는 기술의 변화가 급격하게 변화하고 있으므로, 생성형 AI를 통해서 비식별 대상을 찾고 해당 행위를 인식하며 Diffusion 기술 등을 통한 비식별 처리기의 다양성을 가져오는 연구가 필요하다.

참고문헌

[1] D.-J. Kim and Y.-G. Jeon, “Vision AI-Based Automatic Selective De-Identification,” *Journal of Digital Contents Society*, Vol. 24, No. 4, pp. 725-734, April 2023. http://dx.doi.org/10.9728/dcs.2023.24.4.725

[2] J. Y. Kim, N. S. Jho, and K. Y. Chang, “Trends in Data Privacy Protection Technologies with Enhanced Utilization,” *Electronics and Telecommunications Trends*, Vol. 35, No. 6, pp. 88-96, December 2020. https://doi.org/10.22648/ETRI.2020.J.350609

[3] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint Face Detection and Alignment Using Multi-Task Cascaded Convolutional Networks,” arXiv:1604.02878, April 2016. https://doi.org/10.48550/arXiv.1604.02878

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” arXiv:1801.07698v3, February 2019. https://doi.org/10.48550/arXiv.1801.07698

[5] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, “RetinaFace: Single-Stage Dense Face Localisation in the Wild,” arXiv:1905.00641, May 2019. https://doi.org/10.48550/arXiv.1905.00641

[6] Y. Xu, W. Yan, H. Sun, G. Yang, and J. Luo, “CenterFace: Joint Face Detection and Alignment Using Face as Point,” arXiv:1911.03599, November 2019. https://doi.org/10.48550/arXiv.1911.03599

[7] Labeled Faces in the Wild [Internet]. Available: https://www.kaggle.com/datasets/jessicali9530/lfwdataset

[8] Shuo Yang. WIDER FACE: A Face Detection Benchmark [Internet]. Available: http://shuoyang1213.me/WIDERFACE/.

[9] GitHub. Ultralytics / YOLO Vision [Internet]. Available: https://github.com/ultralytics/yolov5.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single Shot MultiBox Detector,” in *Proceedings of the 14th European Conference on Computer Vision (ECCV 2016)*, Amsterdam, Netherlands, pp. 21-37, October 2016. https://doi.org/10.1007/978-3-319-46448-0_2

[11] R. Girshick, “Fast R-CNN,” in *Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 1440-1448, December 2015. https://doi.org/10.1109/ICCV.2015.169

[12] S. Shinde, A. Kothari, and V. Gupta, “YOLO Based Human Action Recognition and Localization,” *Procedia Computer Science*, Vol. 133, pp. 831-838, 2018. https://doi.org/10.1016/j.procs.2018.07.112

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... and N. Houlsby, “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv:2010.11929v2, June 2021. https://doi.org/10.48550/arXiv.2010.11929

[14] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, “Multiscale Vision Transformers,” in *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, pp. 6804-6815, October 2021. https://doi.org/10.1109/ICCV48922.2021.00675

김대진(Dae-Jin Kim)

1998년 : 대진대학교 대학원 (공학사)
2000년 : 동국대학교 대학원 (공학석사)
2010년 : 대진대학교 대학원 (공학박사)
2017년~현 재: 동국대학교 영상문화콘텐츠연구원 조교수
※관심분야 : 코덱, 멀티미디어 플랫폼, 콘텐츠 DNA, 워터마크, 딥러닝, 번호인식, 주차관제시스템 등

전윤걸(Youn-Girl Jeon)

2004년 : 서울과학기술대학교 메체공학과(공학사)
2014년 : 서울과학기술대학교 대학원 (체육학 석사)
2020년 : 성균관대학교 스포츠과학대학 (체육학 박사)
2020년~현 재: 성균관대학교 체력과학연구소 선임연구원
※관심분야 : 인간동작분석, 자세 평형성, 인공지능, 디지털 헬스케어, 스포츠 경기력, 이상동작 탐지, 노인건강, 인지장애 등

김준화(Jun-Hwa Kim)

2019년 : 동국대학교 (공학사)
2020년 : 동국대학교 대학원 (공학석사)
2023년 : 동국대학교 대학원 (공학박사)
2023년～2024년: 동국대학교 박사후연구원
2024년～현 재: 건양대학교 인공지능학과 조교수
※관심분야 : 영상처리, 컴퓨터비전, 신호처리, 딥러닝 등