ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 March: 28(1): 89~106 http://dx.doi.org/10.13088/jiis.2022.28.1.089

계층적 군집화 기반 Re-ID를 활용한 객체별 행동 및 표정 검출용 영상 분석 시스템

이상현
아주대학교 정보통신대학 소프트웨어학과
(leesh1510@ajou.ac.kr)

양성훈
명지대학교 ICT융합대학 융합소프트웨어학부
(didtjdgns852@gmail.com)

오승진
순천향대학교 의료과학대학 의료IT공학과
(hello641004@gmail.com)

강진범
자이냅스
(jb.kang@xinapse.ai)

최근 영상 데이터의 급증으로 이를 효과적으로 처리하기 위해 객체 탐지 및 추적, 행동 인식, 표정 인식, 재식별(Re-ID)과 같은 다양한 컴퓨터비전 기술에 대한 수요도 급증했다. 그러나 객체 탐지 및 추적 기술은 객체의 영상 촬영 장소 이탈과 재등장, 오클루전(Occlusion) 등과 같이 성능을 저하시키는 많은 어려움을 안고 있다. 이에 따라 객체 탐지 및 추적 모델을 근간으로 하는 행동 및 표정 인식 모델 또한 객체별 데이터 추출에 난항을 겪는다. 또한 다양한 모델을 활용한 딥러닝 아키텍처는 병목과 최적화 부족으로 성능 저하를 겪는다. 본 연구에서는 YOLOv5기반 DeepSORT 객체추적 모델, SlowFast 기반 행동 인식 모델, Torchreid 기반 재식별 모델, 그리고 AWS Rekognition의 표정 인식 모델을 활용한 영상 분석 시스템에 단일 연결 계층적 군집화(Single-linkage Hierarchical Clustering)를 활용한 재식별(Re-ID) 기법과 GPU의 메모리 스루풋(Throughput)을 극대화하는 처리 기법을 적용한 행동 및 표정 검출용 영상 분석 시스템을 제안한다. 본 연구에서 제안한 시스템은 간단한 메트릭을 사용하는 재식별 모델의 성능보다 높은 정확도와 실시간에 가까운 처리 성능을 가지며, 객체의 영상 촬영 장소 이탈과 재등장, 오클루전 등에 의한 추적 실패를 방지하고 영상 내 객체별 행동 및 표정 인식 결과를 동일 객체에 지속적으로 연동하여 영상을 효율적으로 분석할 수 있다.

주제어 : 객체 추적, 재식별, 행동 인식, 표정 인식, 영상 분석

논문접수일 : 2021년 11월 25일 논문수정일 : 2022년 1월 8일 게재확정일 : 2022년 1월 14일
원고유형 : 학술대회용 Fast Track 교신저자 : 강진범

1. 개요

최근에 스마트폰, CCTV, 블랙박스, 고화질 카메라 등으로부터 수집되는 영상 데이터의 양이 급격히 증가했고 이에 따라 비정형 영상 빅데이터를기반으로 사람이나 객체 등을 인식하여 의미있는 정보를 추출하고 내용을 시각적으로 분석하고 활용하기 위한 요구사항이 증대되고 있다(Ko et al., 2014). 또한 많은 산업 분야에서 영상을 분석하기 위한 숙련된 인력의 부족으로 머신러닝과 인공지능을 활용해 영상물을 분석하고 그 결과를 이용해 인력을 보조하려 노력하고 있고(Singh, 2018)(Wright et al., 2020), 그 예로 영상 프레임 분석을 통한 대용량 캡슐내시경 영상의 지능형 판독보조 시스템(Lee et al., 2009), 시각적 특징을 기반한 샷 클러스터링을 통한 비디오 씬 탐지 기법(Shin et al., 2012), 정확히 재가중되는 온라인 전체 에러율 최소화 기반의

객체 추적(JANG et al., 2019) 등의 연구가 수행되었다.

영상 내 객체 별 행동 인식 모델과 표정 인식 모델은 각각 객체 인식 모델과 얼굴 인식 모델을 기반으로 객체 위치와 얼굴 위치의 특징점을 이용하여 구현된다(Herath et al., 2017)(Azizan and Khalid, 2018). 먼저 객체별 위치를 추적하기 SORT(Simple Online and Realtime Tracking)와 여기에 딥러닝 특징(Feature)을 이용하는 2단(Two-step) 방식의 객체 추적 알고리즘 DeepSORT가 등장했다(Bewley et al., 2016)(Wojke et al., 2017). 그러나 DeepSORT에서 실시간 성능을 보장하기 위해 재식별을 위한 메트릭(Metric)으로 단순 최근접 이웃(Simple Nearest Neighbor)을 사용하여 객체의 영상 장소 이탈 및 재등장 상황이나 긴 시간 동안의 오클루전이 발생한 상황에서는 정확도가 현저히 떨어진다.

본 연구는 높은 정확도와 처리 성능을 가지면서 객체별 행동 및 표정 분석을 위해 YOLOv5 (Glenn et al., 2021) 기반 DeepSORT 객체추적 모델, SlowFast(Feichtenhofer et al., 2019) 기반 행동 인식 모델, Torchreid(Zhou and Xiang, 2019) 기반 재식별 모델, 그리고 AWS Rekognition(Amazon, 2021)의 표정 인식 모델을 활용한 영상 분석 시스템에 단일 연결 계층적 군집화(Single-linkage Hierarchical Clustering)를 활용한 재식별 기법을 적용한 영상 분석 시스템을 제안한다. 제시한 연구 모형은 객체의 영상 장소 이탈 후 재등장 또는 오클루전 상황 속에서 추적에 실패한 같은 객체에 대해 재추적이 가능하다. 또한 GPU의 메모리 스루풋을 최대화하면서 램 메모리 요구사항을 낮출 수 있는 객체별 경계 상자 큐(Bounding Box Queue by Object)와 특징 큐(Feature Queue) 기법, AWS Rekognition을 통해 인식한 표정이 객체 추적 정보와 연동될 수 있도록 하는 IoF (Intersection over Face) 알고리즘을 소개한다.

본 연구의 구성은 다음과 같다. 2장에서는 본 연구가 제안하는 시스템에 사용된 객체 탐지 및 추적, 재식별, 행동 인식, 표정 인식과 관련된 컴퓨터비전 기술의 이론적 배경과 선행 연구에 관해 설명한다. 3장에서는 본 연구에서 제안하는 시스템에 관해 설명한다. 4장에서는 실험 데이터와 결과에 대해서 설명한다. 마지막으로 5장에서는 결론에 관해 설명한다.

2. 관련 연구

2.1. 객체 탐지 및 추적(Object Detection & Tracking)

객체 탐지는 디지털 이미지와 영상에서 서로 다른 종류의 객체를 식별하거나, 구분하거나 또는 분류하기 위한 컴퓨터 비전 기술이다. 객체 탐지 기술은 CNN(Convolutional Neural Network)의 개발, 딥러닝 알고리즘의 발전, 그리고 GPU의 컴퓨팅 파워가 증가함에 따라 아주 빠르게 발전하고 있다(Jaioet al., 2019). 이에 따라 객체 탐지 기술은 광범위한 영역에서 활발히 활용되고 있는데, 보안과 감시 시스템에서 목표 객체를 탐지하거나(Wang et al., 2017) (Bashir et al., 2019) 의료과학 시스템에서 피부암을 진료하는(Younis et al., 2019) 등 사람보다 나은 정확도로 다양한 문제를 해결하는 데 성공했다.

딥러닝을 기반으로 한 객체 탐지는 크게 2단계(Two-stage) 방식과 1단계(One-stage) 방식으로 나눌 수 있다. 2단계 방식은 먼저 이미지에서 객체의 위치를 지정하고, 그 다음 분류기가 해당

위치에 있는 객체를 분류하고, 1단계 방식은 객체의 위치 지정과 분류를 한 번에 수행한다. 과거에는 1단계 방식의 객체 탐지가 실시간 성능을 보장했으나 정확도 측면에서 2단계 방식의 것보다 낮았는데, 최신 딥러닝 기법을 적용한 YOLOv4(Bochkovskiy et al., 2020)와 같은 모델이 등장해 높은 정확도를 보였다(Gudelj et al., 2021). 최근에는 YOLOv5 모델이 소개되었는데, YOLOv5는 YOLOv4와 동일하게 CSP와 PA-NET을 사용했지만 모자이크 데이터 증강(Mosaic Data Augmentation)과 경계 상자 앵커 자동 학습(Auto Learning Bounding Box Anchors) 기법을 적용한 것이 차이점이다(Glenn et al., 2021).

다중 객체 추적(Multi Object Tracking, MOT)은 영상 내에서 보행자, 자동차, 동물 등의 여러 객체들을 사전 정보 없이 식별하고 추적하기 위한 기술이다(Ciaparrone et al., 2019). 딥러닝의 발전에 따라 객체의 특징을 추출해 딥러닝 기반의 재식별 모델을 이용하는 DeepSORT 알고리즘이 등장했다(Wojke et al., 2017). 오늘날 객체 추적 알고리즘의 대다수가 DeepSORT 알고리즘의 구성 요소를 포함하거나 일부 공유한다. 이 때 객체 탐지 및 추적과 재식별 기능을 개별적으로 수행하는 방식을 2단 방식, 그렇지 않은 경우가 원샷 방식으로 최근에는 원샷 방식의 알고리즘이 활발히 연구되고 있다(Zhang et al., 2020).

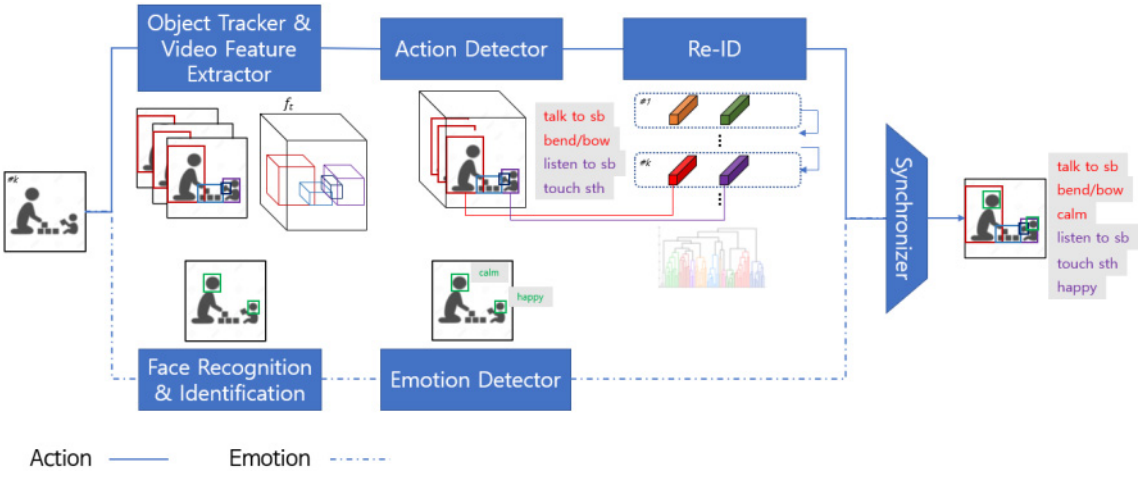
2.2. 사람 재식별(Person Re-Identification)

사람 재식별 기술의 주 목표는 서로 다른 카메라 상에서 같은 인물을 식별하는 것이다. 오늘날 재식별 기술은 크게 시점 변화에 강인한 특징 설계와, 특징 매핑 및 거리 학습기법으로 분류할 수 있다(An et al., 2017)(Ye et al., 2021). 첫 번째는 시점과 환경 변화에 강인한 특징 설명자(Feature Descriptor)를 고안하는 것을 목표로 가지며, 예로 조명에 영향받지 않는 색 설명자(Color Descriptors)(Kviatkovsky et al., 2013), 의미론적 색 이름(Semantic Color Names)(Kuo et al., 2013), 국소 최대 발생 특징(Local Maximal Occurrence Feature)(Liao et al., 2015), 참조 설명자(Reference Descriptor)(An et al., 2016) 등이 있다. 두 번째는 다른 환경의 같은 객체의 특징 간 차이를 효과적으로 줄일 수 있는 특징 매핑법과 거리 메트릭을 학습하는 방법을 고안하는 것을 목표로 한다. 예시로 상대 거리 비교(Relative Distance Comparison)(Zheng et al., 2013), 국소 피셔법에 의한 선형 판별 분석(Local Fisher Discriminant Analysis)(Pedagadi et al., 2013), 강인한 정준형 상관 분석(Robust Canonical Correlation Analysis)(An et al., 2015) 등이 있다.

딥러닝에 대한 관심과 발전이 지속됨에 따라 딥러닝과 관련된 오픈소스 프레임워크에 대한 연구도 활발히 진행되고 있다. 대표적으로 Caffe(Jiaet al., 2014), PyTorch(Paszke et al., 2017), TensorFlow (Abadi et al., 2016), MXNet(Chen et al., 2015) 등이 있다. 그러나 이들은 서로 다른 백엔드 프레임워크, 데이터 처리 모듈, 평가 절차를 가지고 있어 표준화된 인터페이스를 갖지 않는다. 이런 문제로 딥러닝 기반 재식별 모델 개발에 차질이 생기자 이를 위한 범용 프레임워크 Torchreid가 등장했다(Zhou et al., 2019).

2.3. 행동 탐지(Action Detection)

영상의 행동 이해 기술은 하나 혹은 그 이상의 행동을 포함하도록 분할 편집된 비디오(Well-trimmed video)에 대해 그 비디오가 표현하는 행동 클래

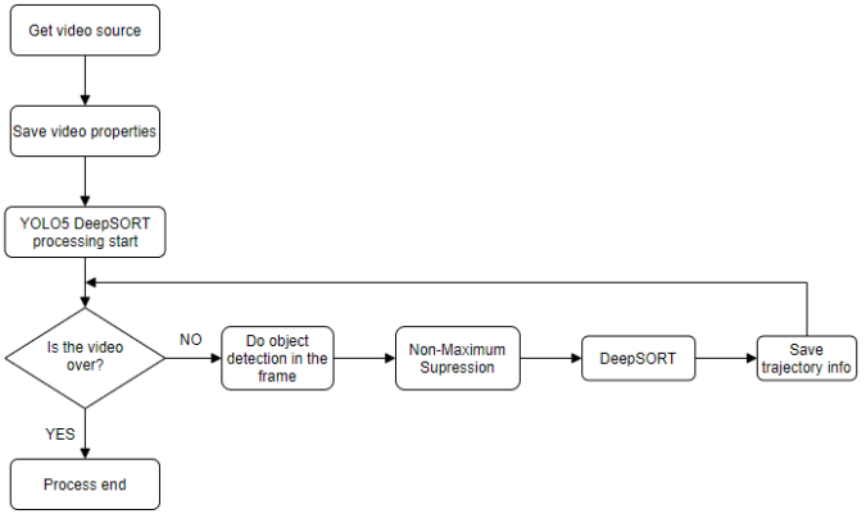


〈Figure 1〉 Overall Process Architecture

스를 분류하는 행동 인식(Action Recognition) 기술과, 행동을 포함하지 않는 백그라운드와 여러 행동 클래스에 속하는 다수의 행동 인스턴스들을 포함하는 일반 무편집 비디오(Un-trimmed Video)에서 각 행동 인스턴스별로 행동의 발생 위치를 추정하고 행동을 인식하는 행동 탐지(Action Detection) 기술로 구분된다. 실세계 비디오에서 행동을 이해하기 위해서는 행동 탐지 기술이 필수적이다. 행동 탐지 기술 중에서 위치 추정의 대상이 시간에 한정된 경우를 시간적 행동 탐지(Temporal Action Detection) 기술이라고 하고, 시공간에 대해서, 즉 시간 구간에 포함된 각 프레임에서의 행동이 발생한 공간적 위치까지 추정하는 경우를 시공간 행동 탐지(Spatio-Temporal Action Detection) 기술이라고 한다. 오늘날 대부분의 행동 탐지 연구들은 주로 현실적인 시간적 행동 탐지 기술에 대해 진행되고 있다(Moon et al., 2020).

2.4. 표정 탐지(Emotion Detection)

얼굴에 나타나는 표정은 인간 감정을 파악하는 데 핵심적인 요소 중 하나로, 인간이 감정을 드러내는 데 얼굴 표정과 같은 비언어적 부분이 55%를 차지하며 얼굴 표정은 진심을 전달하고 있다는 확실한 증거이다(Bartlett et al., 2008). 최근에는 로봇과 소셜 네트워킹 등 다양한 산업 분야에서 더 나은 사용자 경험을 위해 사람의 감정을 파악하려는 시도가 이루어지고 있는데, 사람은 비교적 간단히 얼굴 표정을 탐지하는 데에 반해 기계는 많은 단계를 거쳐야 한다. 표정을 탐지하기 위해선 먼저 얼굴의 위치를 특정한 다음 특정된 얼굴 위치로부터 특징 벡터를 추출한다. 이후 추출된 특징 벡터를 이용해 표정을 분류한다. 얼굴 표정을 탐지하는 자동화된 모델을 가지기 위해선 얼굴 표정 데이터들을 가진 데이터베이스가 필요하다. 특징 추출기가 얼굴 이미지들로부터 특징 벡터를 추출하고 추출한 데이터들을 표정에 따라 데이터베이스에 임베딩(Embedding)한다. 이후에 새 이미지에 대한 표정을 쿼리하면 새 이미지의 얼굴 표정 특징 벡터를 추출하고 이를 임베딩된 벡터들과 비교하여 가장 유사한 표정으로 분류한다



〈Figure 2〉 Object Tracker Process Flow

(Moolchandani et al., 2021).

3. 시스템 설계

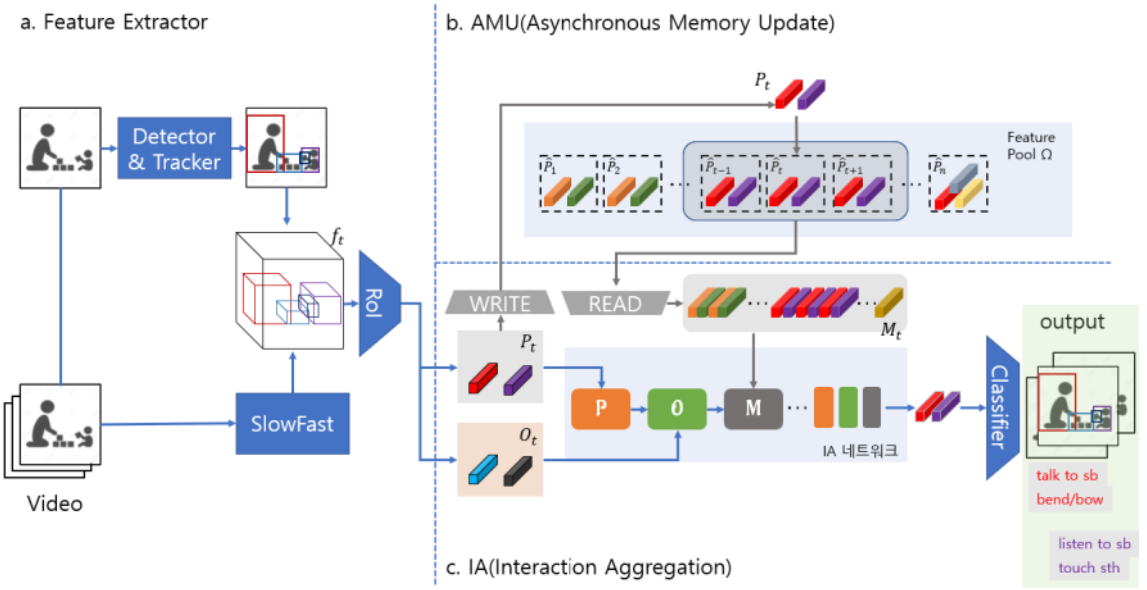
본 연구에서는 영상을 입력받으면 영상 내 객체별 행동 및 표정을 검출하고 메타데이터를 추출·시각화하는 기능을 수행하는 영상 분석 시스템을 제안한다. <Figure 1>은 본 연구에서 제안하는 시스템의 전체 프로세스에 대한 아키텍처이다. 시스템은 크게 두 가지 트랙으로 병렬 처리된다. 하나는 객체의 행동을 검출하기 위한 행동 검출 프로세스, 다른 하나는 객체의 표정을 검출하기 위한 표정 검출 프로세스이다.

3.1. Object Tracker

영상 내 등장 객체들의 위치를 좌표 값으로 나타내고, 해당 객체의 과거 정보를 바탕으로 움직임을 예측·추적하는 기능을 수행한다. 영상 프레임마다 YOLOv5 모델로 객체를 탐지하면 객체별로 위치를 예측하는 여러 경계 상자가 생기는데, 이 중 신뢰도가 가장 높은 상자를 선택하고 해당 상자와의 IoU(Intersection of Union) 임계치보다 이상이면 제외(Suppression)한다. 이후 DeepSORT 알고리즘을 이용하여 객체를 추적한다. 이 과정에 대한 순서도를 <Figure 2>에서 확인할 수 있다.

3.2. Video Feature Extractor & Action Detector

시공간 행동 탐지 기능을 수행하기 위해 SlowFast 망(Feichtenhofer et al., 2019)을 중추망으로 한 AIA(Asynchronous Interaction Aggregation for Action Detection)(Tang et al., 2020) 모델을 사용한다. AIA 모델에 대한 아키텍처를 <Figure 3>에서 확인할 수 있다. 해당 모델은 먼저 영상을 64 프레임 단위 여러 개의 클립으로 나누고, 클립으로부터 SlowFast 망이 비디오 특징을 추출한 뒤

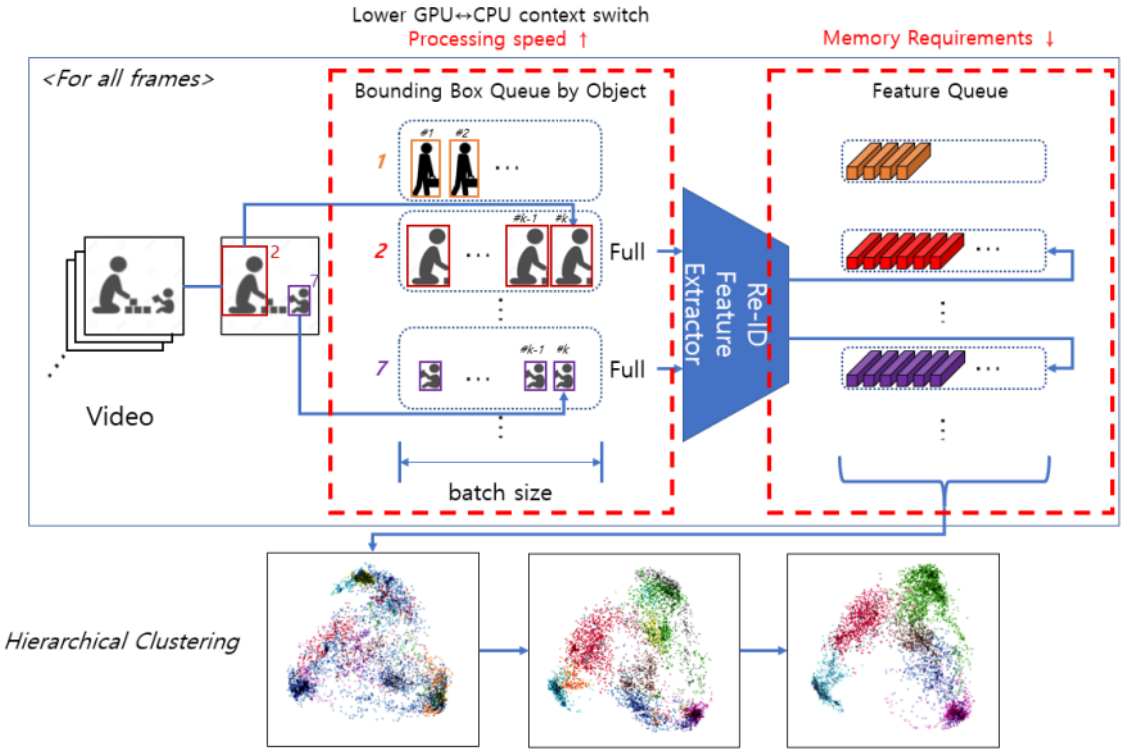


〈Figure 3〉 Video Action Detection Process Architecture

객체 추적 모델이 탐지한 객체별 위치 정보를 기반으로 객체별 특징점을 정렬한다(RoI Align). 이후 해당 특징 벡터를 특징 풀(Feature Pool)에 적재하고, 과거의 특징 벡터 비교를 통해 비디오 내 객체의 시간 정보를 업데이트한다(Asynchronous Memory Update). 최종적으로 시간 정보를 담고 있는 메모리 특징 벡터와 공간 정보를 담고 있는 사람 객체의 특징 벡터, 사물의 특징 벡터를 상호작용 통합(Interaction Aggregation) 딥러닝 망에 입력하여 행동을 분류한다.

3.3. Re-ID

시스템 내 재식별 모델이 수행하는 기능과 그 순서를 <Figure 4>에서 확인할 수 있다. 재식별 모델은 객체 추적 모델을 통해 얻은 객체별 경계 상자 이미지에서 재식별 특징을 추출한 뒤 과거 프레임에 등장했던 객체의 특징들과 비교하여 계층적 군집화를 진행한다. 각 객체의 경계 상자 이미지로부터 추출한 재식별 특징 벡터들은 유클리드 공간에 임베딩되는데, 임베딩이 모두 끝나면 먼저 등장했던 객체와 새로 등장한, 즉 새로운 ID를 가지는 각 객체의 특징 벡터 군집과 단일 연결 계층적 군집화를 수행하여 추적에 실패했던 객체를 과거 동일 객체와 연동하고 재추적한다. 이를 통해 객체가 영상 밖으로 나간 후 다시 등장한 경우나 객체 혹은 사물에 의해 가려진 후 다시 등장한 경우 등 추적에 실패한 상황에서 동일 객체를 재식별하고 추적을 지속할 수 있도록 한다. 한 편, 연구 모형에 오픈소스로 활용할 수 있는 재식별 모델(samihormi, 2020)을 바로 적용했을 때 처리 속도가 느리고 메모리 요구사항이 높은 문제점이 있었다. 본 연구에서는 이를 해결하기 위해 객체별 경계 상자 큐(Bounding Box Queue by Object)와 특징 큐(Feature Queue)를 설계했다.



〈Figure 4〉 Re-ID Process Architecture

3.3.1. Bounding Box Queue by Object

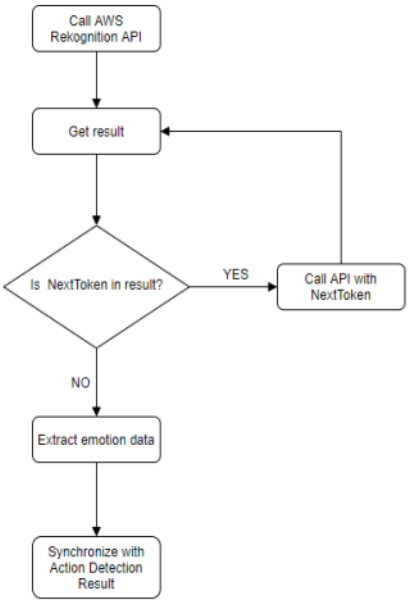
객체별 경계 상자 이미지 한 개씩 재식별 특징을 추출하는 경우 CPU-GPU 간 문맥 교환(Context-switch) 빈도가 높아 처리 속도가 몹시 저하되었다. 이를 해결하기 위해 객체별 이미지를 큐에 쌓아두어 분석 시스템의 GPU 메모리 크기에 맞게 배치 단위로 특징을 추출한다. 이를 통해 문맥 교환 빈도를 최소화함으로써 처리 속도가 배치 사이즈에 선형적으로 증가한다.

3.3.2. Feature Queue

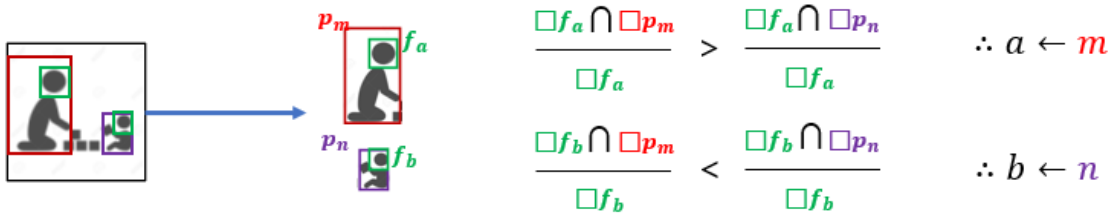
객체 추적을 수행하면서 객체의 경계 상자 이미지들을 모두 모은 뒤 재식별 특징을 추출하도록 설계된 경우 매우 큰 해상도의 원본 이미지 데이터를 보관하기 때문에 비효율적이다. 이를 해결하기 위해 프레임에 따라 객체별 경계 상자 이미지에서 특징을 추출한 뒤, 더 이상 사용하지 않는 이미지들은 보관하지 않고 버리면서 추출한 특징은 객체별 큐에 적재한다. 이를 통해 시스템 메모리를 효율적으로 관리하면서 프레임에 따라 추출된 순서를 보존한다.

3.4. Face-Emotion Detector & Synchronizer

본 연구에서는 영상 내 표정 검출을 위해 AWS Rekognition(Amazon, 2021)의 표정 인식 모델을 활용하였다. 시스템이 AWS Rekognition을



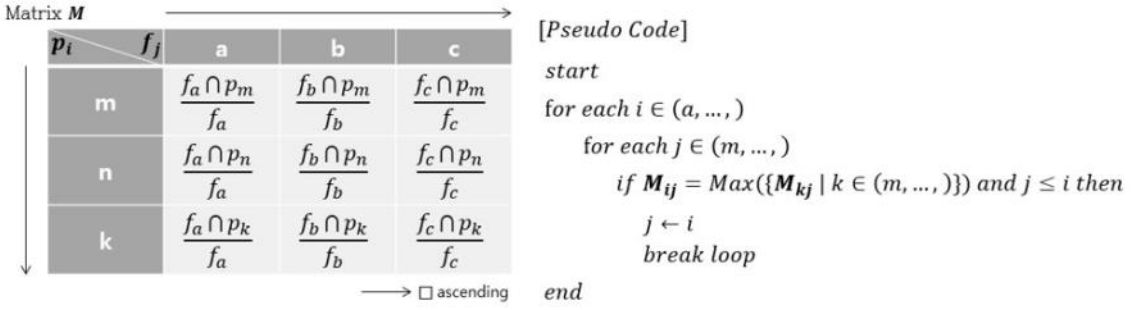
〈Figure 5〉 Emotion Detection & Synchronization Process Flow



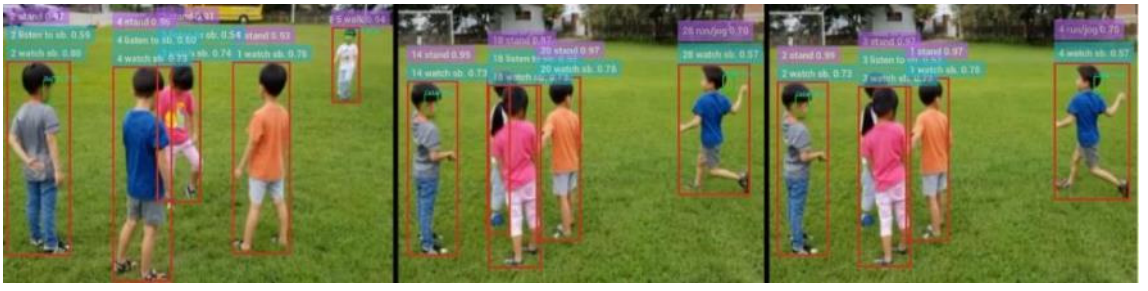
〈Figure 6〉 Basic IoF algorithm

이용하는 방법은 <Figure 5>의 순서도와 같다. 시스템 내부에서 AWS Rekognition API를 호출하면 영상의 일정 프레임마다 얼굴을 인식하고 얼굴 특징을 추출한 뒤 표정을 검출한다. 이후 검출한 표정을 행동 검출 프로세스에서 얻은 객체별 ID에 연동하여 객체별로 시간에 따라 어떤 표정을 지녔었는지 알 수 있다. 객체별 ID에 연동할 때 본 연구에서는 <Figure 6>의 IoF(Intersection over Face) 알고리즘을 사용했다.

영상 프레임에서 검출한 얼굴의 경계 상자와 객체별 경계 상자 면적을 바탕으로 <Figure 6>과 같은 수식을 통해 면적 비율 IoF를 구하여 IoF 값이 가장 큰 객체에 표정 검출 결과를 할당한다. <Figure 6> 수식만으로는 객체가 겹쳐있을 경우 겹친 모든 객체에 대한 IoF 값이 1로 동일하므로 원하는 대로 동작하지 않을 수 있다. 이 문제를 해결하기 위해 <Figure 7>의 향상된 IoF 알고리즘이 필요하다. 가장 큰 IoF 값을 가지는 객체가 여럿일 경우 영상 내 원근에 따라 적절한 객체 ID에 표정을 할당할 수 있도록 탐욕법



〈Figure 7〉 Advanced IoF algorithm



〈Figure 8〉 Test Video Capture: Early part, with no re-id, with re-id

(Greedy Algorithm)을 2차원 행렬 자료구조를 통해 구현하여 약점을 보완한다.

4. 실험 결과

제안된 시스템의 성능과 정확도를 측정하기 위한 실험 데이터로 객체의 영상 촬영 장소 이탈 및 재등장, 객체 간 오클루전이 잦은 영상(중앙육아종합지원센터, 2020)과 부모와 아이의 놀이 상담 영상(KT - 케이티, 2021)을 이용하였다. 영상 분석 시 소요 시간 비교와 분석 결과를 시각적으로 합성한 출력 영상을 기반으로 시스템의 성능과 정확도를 알 수 있도록 실험을 진행했다. <Figure 8>은 테스트 영상에서 객체들이 등장하는 초반부, 재식별 기능을 수행하지 않고 영상을 분석한 결과, 재식별 기능을 정상적으로 수행하여 영상을 분석한 결과를 캡처하여 순서대로 나열한 것이다. 영상 초반부에 등장한 객체들에게 1부터 5까지의 ID가 할당된 모습을 볼 수 있다. 또한 재식별 기능이 없을 때 객체 추적에 실패해 다른 ID들을 부여한 반면 재식별 기능을 수행했을 때 재추적하여 동일한 ID를 부여한 것을 확인할 수 있다. 결과적으로 기존의 행동 및 표정 인식 모델에 재식별 모델을 활용하여 추적 실패를 만회하고 객체별 행동 및 표정 검출 결과를 동일 객체에 적절히 연동할 수 있음을 확인할 수 있다. <Figure 9>은 해당 영상 객체별로 군집화된 재식별 특징을 Tensorflow Projector로 시각화한 것이다. 영상에 등장하는 객체 7명의 재식

〈Figure 9〉 Embedding Feature Visualization(Tensorflow Projector)

〈Table 1〉 Processing time comparison on test video(1 minute length)

	with no Re-ID	Existing Re-ID open-source	Proposed
Action Detection Processing Time(sec)	78	180(78+102)	86(78+8)

〈Table 2〉 Memory requirements comparison on test video

RAM size	Existing Re-ID open-source	Proposed
Per 1 second	30 MB	0.5 MB
Per 1 minute	2 GB	30 MB

별 특징점 군집이 서로 다른 색으로 표현되어 있다.

또한 본 연구에서 제안한 객체별 경계 상자 큐와 특징 큐 기법이 성능을 얼마나 향상시켰는지 측정하기 위해 테스트 영상의 처리 시간과 실제 메모리 사용량을 확인한 결과를 <Table 1>과 <Table 2>를 통해 확인할 수 있다. <Table 1>은 객체별 경계 상자 큐를 활용한 경우 객체별 행동 검출 프로세스에서 10배 이상 처리 속도를 개선했음을 알 수 있다. <Table 2>는 특징 큐를 활용한 경우 오픈소스의 재식별 모델 구조를 그대로 이용한 것보다 메모리 요구사항을 대폭 감소시켰음을 알 수 있다.

마지막으로 IoF 기본 알고리즘이 적절히 동작하는지 여부와 향상된 IoF 알고리즘이 해결하고자 했던 문제 상황을 <Figure 10>를통해 확인할 수 있다. <Figure 10>을 보면 검출된 두 객체와 두 얼굴의 경계 상자로부터 계산한 IoF 계산 값이 모두 1로 동일할 것임을 추측할 수 있다. 이 경우 향상된 IoF 알고리즘을 이용해 객체 경계 상자와 얼굴 경계 상자를 크기 순으로 매핑하여, 인식한 얼굴 위치와 표정을 적절한 객체에 할당한다. 그 결과 화면 좌측에 검출된 표정이 8번 객체에, 우측에 검출된 표정이 1번 객체에

〈Figure 10〉 Situation where IoF advanced algorithm is needed

할당된다.

5. 결론

5.1. 연구 결과 토의

본 연구에서는 단일 연결 계층적 군집화 기반 재식별 기법을 활용하여 영상 내 객체별 행동과 표정을 분석할 수 있는 영상 분석 시스템을 제안했다. 기존 2단 방식의 재식별 모델을 영상 처리에 사용할 때 처리 성능이 저하되어 최대한 메트릭을 단순히 설계했던 방식과 달리 고비용의 계층적 군집화 기법을 사용하여 정확도를 높이고, 객체별 경계 상자 큐 특징 큐와 기법을 이용하여 처리 성능을 향상시켰다. 이를 통해 재등장 및 오클루전 상황에서도 추적 실패로 인한 행동 및 표정 검출 결과가 다른 객체의 것으로 오인식하는 문제를 해결할 수 있으며 GPU의 메모리 스루풋을 극대화하여 처리 성능을 10배 이상 향상시키고 메모리 요구 사항을 대폭 낮출 수 있음을 확인할 수 있었다. 또한 본 연구에서 각각 독자적인 행동 탐지 모델과 표정 탐지 모델이 검출한 결과를 적절히 연동하기 위해 IoF 알고리즘을 소개했다. 이를 통해 객체 ID 정보가 없는 표정 검출 결과를 객체 추적 모델 및 재식별 모델이 산출해낸 ID에 할당하여 객체별 행동과 표정을 동시에 분석할 수 있음을 확인했다.

5.2. 학술적 의의와 실무적 시사점

본 연구의 학술적 의의는 2단 방식의 재식별 모델이 실시간 성능을 갖추기 위해 단순한 메트릭을 사용하여 정확도를 저하시킬 필요 없이 처리 기법에 따라 행동 및 표정 인식을 수행하는 고비용 환경에서도실시간 성능을 갖출 수 있는 가능성을 보였다는 점에 있다. 다양한 영역에서 객체 추적과 재식별 기술을 동반하는 많은 연구가 반비례 관계에 있는 정확도와 처리 성능을 두고 여러 가지 방법을 제시하고 있는데, 본 연구

에서 제안한 모형은 향후 관련 연구에서 정확도와 처리 성능을 모두 높일 수 있는 방법으로서 기여할 수 있다.

본 연구의 실무적 시사점은 영상 내 행동과 표정 분석이 필요한 다양한 산업 분야에서 기존에는 객체가 영상 장소에서 이탈하거나 다른 객체에 의해 빈번히 오클루전이 발생하여 객체별 지속 추적 및 분석이 어려운 환경이더라도 본 연구의 모형을 통해 효과적으로 분석이 가능하다는 점에 있다. 높은 정확도의 재추적 성능과 빠른 처리 성능을 가지는 본 연구 모형이 지능형 감시, 관찰 서비스와 행동 또는 심리 분석 서비스 등 객체별 추적 정보와 영상 분석 메타 데이터의 통합이 산업적, 비즈니스적으로 큰 가치를 창출하는 다양한 분야에서 유용하게 사용될 수 있을 것이다.

5.3. 연구의 한계점 및 향후 계획

본 연구의 재식별 기법을 활용한 객체 추적 성능에 대해 다양한 테스트 데이터셋 검증이 추가적으로 필요하다. 본 연구의 실험은 제시했던 문제 상황을 해결할 수 있음을 시사하는 것에는 부족함이 없지만 성능 향상 정도와 다양한 상황에서의 종합적인 성능을 정량적으로 측정하기 위해 보다 더 다양한 데이터셋에서 규칙을 정하여 테스트를 진행할 필요가 있다. 또한 본 연구의 IoF 알고리즘이 해결하고자 했던 객체별 행동과 표정 검출 결과 연동 문제는 영상 내에서 등장 객체 수보다 적은 수의 표정이 검출되는 경우 표정 할당이 원하는 대로 동작하지 않을 가능성이 있다. 따라서 향후에는 객체 추적 성능을 측정하기 위해 많은 국제 학회에서 사용하는 데이터인 MOT Challenge 데이터셋을 활용한 실험을 진행하고, IoF 알고리즘이 해결하지 못할 것으로 예상되는 문제를 조사하여 그를 해결하기 위한 추가 보완 알고리즘을 적용할 계획이다. 더불어 지능형 영상 분석과 관련된 다양한 분야의 데이터셋에 본 모형을 적용해보는 추가 연구를 진행할 계획이다.

참고문헌(References)

Abadi M, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467*, 2016.

Amazon, "Amazon Rekognition," *AWS*, Available at https://aws.amazon.com/rekognition/ (Accessed Sep, 2021).

An L, S. Yang, and B. Bhanu, "Person re-identification by robust canonical correlation analysis," *IEEE Signal Processing Letters*, Vol.22, No.8 (2015), 1103~1107.

An L., M. Kafai, S. Yang, and B. Bhanu, "Person re-identification with reference descriptor," *IEEE Transactions on Circuits and Systems for Video Technology, Vol.26, No.4 (2016)*, 776~787.

An, L., X. Chen, S. Liu, Y. Lei, and S. Yang, "Integrating appearance features and soft biometrics for person re-identification," *Multimedia Tools and Applications: An International Journal*, Vol.76, No.9 (2017), 12117~12131.

Azizan I., and F. Khalid, "Facial Emotion Recognition: A Brief Review", *International Conference on Sustainable Engineering, Technology*

and Management(ICSETM), 2018.

Bartlett M. S., G. Littlewort, E. Vural, K. Lee, M. Cetin, A. Ercil, and J. Movellan, "Data mining spontaneous facial behavior with automatic expression coding," *Lecture Notes in Computer Science*, Vol.5042, 2008, 1~20.

Bashir M., E. A. Rundensteiner, and R. Ahsan, "A deep learning approach to trespassing detection using video surveillance data," *IEEE International Conference on Big Data*, 2019, 3535~3544.

Bewley A., Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," *IEEE International Conference on Image Processing(ICIP)*, 2016, 3464~3468.

Bochkovskiy A., C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv:2004.10934 [cs.CV]*, 2020.

Broström M., "Real-time multi-object tracker using YOLOv5 and deep sort,", *Github*, 2021, Available at https://github.com/mikel-brostrom/Yolov5_DeepSort_Pytorch/ (Accessed Sep, 2021).

Chen T., M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv:1512.01274*, 2015.

Ciaparrone G., F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep Learning in Video Multi-Object Tracking: A Survey," *arXiv:1907.12740*, 2019.

Feichtenhofer C., H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition", *IEEE/CVF International Conference on Computer Vision(ICCV)*, 2019, 6201~6210.

Glenn J. et al., (2021). "ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support (v6.0)," *Zenodo*, 2021, Available at https://doi.org/10.5281/zenodo.5563715/ (Accessed Sep, 2021).

Gudelj D., A. F. Stama, J. Petroviæ and P. Pale, "Visual Object Detection - an Overview of Algorithms and Results," *44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2021, 1727~1732.

Herath S., M. Harandi, and F. Porikli, "Going Deeper into action recognition: A survey", *Image and Vision Computing*, Vol.60, No.4 (2017), 4~21.

Jaio L. et al., "A Survey of Deep Learning-Based Object Detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* vol.7 (2019), 128837~128868.

JANG, S.-I., and C.-S. . PARK, "Object Tracking Based on Exactly Reweighted Online Total-Error-Rate Minimization", *Journal of Intelligence and Information Systems*, Vol. 25, No. 4 (2019), 53~65.

Jia Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv:1408.5093*, 2014.

Ko J. G., Y. S. Bae, J.Y. Park, and K. Park, "Technologies Trends in Image Big Data Analysis," *Electronics and Telecommunications Research Institute(ETRI)*, Vol.29, No.4 (2014), 21~29.

KT - 케이티, “8년 경력 어린이집 교사 엄마도 반한 우리 아이 현실 육아 비법은?! [가정교사] Ep.4,” *Youtube*, Jan. 6, 2021, Availble at

https://www.youtube.com/watch?v=jS2e8iKAqP4 (Accessed Sep, 2021)

Kuo C. -H., S. Khamis, and V. Shet "Person re-identification using semantic color names and rankboost," *IEEE Workshop on applications of computer vision(WACV)*, 2013, 281~287.

Kviatkovsky I, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.35, No.7 (2013), 1622~1634.

Lee H. G., M. K. Choi, D. H. Lee, and S. C. Lee, "Intelligent Diagnosis Assistant System of Capsule Endoscopy Video Through Analysis of Video Frames", *Journal of Intelligence and Information Systems*, Vol. 15, No. 2 (2009), 33~48.

Liao S, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015, 2197~2206.

Moolchandani M., S. Dwivedi, S. Nigam and K. Gupta, "A survey on: Facial Emotion Recognition and Classification," *5th International Conference on Computing Methodologies and Communication(ICCMC)*, 2021, 1677~1686.

Moon J. Y., H. I. Kim, and J. Y. Park, "Trends in Temporal Action Detection in Untrimmed Videos," *Electronics and Telecommunications Research Institute(ETRI)*, Vol.35, No.3 (2020), 20~33.

Paszke A., S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *NIPS*, 2017.

Pedagadi S, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, 3318~3325.

samihormi, "Multi-Camera-Person-Tracking-and-Re-Identification," *Github*, 2020, Available at https://github.com/samihormi/Multi-Camera-Person-Tracking-and-Re-Identification (Accessed Jul, 2021).

Shin, D.-W., T.-H. Kim, and J.-M. Choi, "Video Scene Detection using Shot Clustering based on Visual Features", *Journal of Intelligence and Information Systems*, Vol. 18, No. 2 (2012), 47~60.

Singh B., "DETECTING OBJECTS AND ACTIONS WITH DEEP LEARNING," (PhD Thesis), *University of Maryland, College Park*, 2018.

Tang J., J. Xia, X. Mu, B. Pang, and C. Lu, "Asynchronous Interaction Aggregation for Action Detection," *IEEE Conference on Computer Vision and Pattern Recognition(CVPR),* 2020.

Wang Y., T. Bao, C. Ding, and M. Zhu, "Face recognition in real-world surveillance videos with deep learning method," *2nd International Conference on Image, Vision and Computing (ICIVC)*, 2017, 239~243.

Wojke N., A. Bewley, and D. Paulus, "Simple oneline and realtime tracking with a deep association metric," *IEEE International Conference on Image Processing(ICIP)*, 2017, 3645~3649.

Wright C. et al., "AI IN PRODUCTION: VIDEO ANALYSIS AND MACHINE LEARNING FOR EXPANDED LIVE EVENTS COVERAGE," *SMPTE Motion Imaging Journal*, vol.129, No.2 (2020), 36~45.

Ye M., J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep Learning for Person Re-identification: A Survey and Outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Younis H., M. H. Bhatti, and M. Azeem, "Classification of Skin Cancer Dermoscopy Images using Transfer Learning," *15th International Conference on Emerging Technologies(ICET)*, 2019, 1~4.

Zhang Y., C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the Fairness of Detection and Re-identification Object Tracking", *arXiv:2004.01888*, 2020.

Zheng W., S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.35, No.3(2013), 653~668.

Zhou K. and T. Xiang, "Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch," *arXiv:1910.10093 [cs.CV]*, 2019.

중앙육아종합지원센터, “2019 개정 누리과정 시범어린이집 놀이영상(전궁몬테소리어린이집),” *Youtube*, Available at https://www.youtube.com/watch?v=P8P0ZMP4nZo/ (Accessed 27 Mar, 2020).

Abstract

Video Analysis System for Action and Emotion Detection by Object with Hierarchical Clustering based Re-ID

Sang-Hyun Lee* · Seong-Hun Yang** · Seung-Jin Oh*** · Jinbeom Kang****

Recently, the amount of video data collected from smartphones, CCTVs, black boxes, and high-definition cameras has increased rapidly. According to the increasing video data, the requirements for analysis and utilization are increasing. Due to the lack of skilled manpower to analyze videos in many industries, machine learning and artificial intelligence are actively used to assist manpower. In this situation, the demand for various computer vision technologies such as object detection and tracking, action detection, emotion detection, and Re-ID also increased rapidly. However, the object detection and tracking technology has many difficulties that degrade performance, such as re-appearance after the object's departure from the video recording location, and occlusion. Accordingly, action and emotion detection models based on object detection and tracking models also have difficulties in extracting data for each object. In addition, deep learning architectures consist of various models suffer from performance degradation due to bottlenects and lack of optimization.

In this study, we propose an video analysis system consists of YOLOv5 based DeepSORT object tracking model, SlowFast based action recognition model, Torchreid based Re-ID model, and AWS Rekognition which is emotion recognition service. Proposed model uses single-linkage hierarchical clustering based Re-ID and some processing method which maximize hardware throughput. It has higher accuracy than the performance of the re-identification model using simple metrics, near real-time processing performance, and prevents tracking failure due to object departure and re-emergence, occlusion, etc. By continuously linking the action and facial emotion detection results of each object to the same object, it is possible to efficiently analyze videos.

* Department of Software and Computer Engineering, Ajou University
** Department of Convergence Software, Myongji University
*** Department of Medical Information Technology Engineering, Soonchunhyang University
**** Corresponding author: Jinbeom Kang
Chief Technology Officer, Xinapse
14 Teheran-ro 86-gil, Gangnam-gu, Seoul (06179), Korea
Tel: *** - **** - **** E-mail: jb.kang@xinapse.ai

The re-identification model extracts a feature vector from the bounding box of object image detected by the object tracking model for each frame, and applies the single-linkage hierarchical clustering from the past frame using the extracted feature vectors to identify the same object that failed to track. Through the above process, it is possible to re-track the same object that has failed to tracking in the case of re-appearance or occlusion after leaving the video location. As a result, action and facial emotion detection results of the newly recognized object due to the tracking fails can be linked to those of the object that appeared in the past. On the other hand, as a way to improve processing performance, we introduce Bounding Box Queue by Object and Feature Queue method that can reduce RAM memory requirements while maximizing GPU memory throughput. Also we introduce the IoF(Intersection over Face) algorithm that allows facial emotion recognized through AWS Rekognition to be linked with object tracking information.

The academic significance of this study is that the two-stage re-identification model can have real-time performance even in a high-cost environment that performs action and facial emotion detection according to processing techniques without reducing the accuracy by using simple metrics to achieve real-time performance. The practical implication of this study is that in various industrial fields that require action and facial emotion detection but have many difficulties due to the fails in object tracking can analyze videos effectively through proposed model. Proposed model which has high accuracy of retrace and processing performance can be used in various fields such as intelligent monitoring, observation services and behavioral or psychological analysis services where the integration of tracking information and extracted metadata creates greate industrial and business value.

In the future, in order to measure the object tracking performance more precisely, there is a need to conduct an experiment using the MOT Challenge dataset, which is data used by many international conferences. We will investigate the problem that the IoF algorithm cannot solve to develop an additional complementary algorithm. In addition, we plan to conduct additional research to apply this model to various fields' dataset related to intelligent video analysis.

Key Words : Object Detection, Re-identification, Action Detection, Emotion Detection, Video Analysis

Received : November 25, 2021 Revised : January 8, 2022 Accepted : January 14, 2022
Corresponding Author : Jinbeom Kang

저 자 소 개

이 상 현

아주대학교 정보통신대학 소프트웨어학과 학생으로 재학중이다. 과학기술정보통신부에서 주관하는 창의도전형 SW인재 육성 프로그램 SW 마에스트로 12기를 수료하였다. 주요 관심 분야는 Object Detection, Multi-Object Tracking, Re-Identification, Spatio-Temporal Action Detection, Facial-Emotion Detection, Feature Embedding, Metric Learning, Explainable AI 등이다.

양 성 훈

현재 명지대학교 융합소프트웨어학부 데이터테크놀로지전공 학생으로 재학중이다. 과학기술정보통신부에서 주관하는 창의도전형 SW인재 육성 프로그램 SW마에스트로 12기를 수료하였다. 주요 관심 분야는Web Framework, Clean Architecture, SPA(Single-Page Application), HCI, Usability Test, UI/UX, Software Quality 등이다.

오 승 진

순천향대학교 의료IT공학과와 자기주도설계전공 회계빅데이터학과 학생으로 재학중이다. 과학기술정보통신부에서 주관하는 창의도전형 SW인재 육성 프로그램 SW 마에스트로 12기를 수료하였다. 주요 관심 분야는 Cloud Architecturing, Application Clustering, Performance Tuning, SW Engineering 등이다.

강 진 범

현재 인공지능 스타트업 자이냅스(Xinapse)에서 CTO로 재직하고 있다. 한양대학교 컴퓨터공학으로 공학석사와 박사학위를 취득하였다. LG전자 MC사업본부, 신한은행 AI랩에서 인공지능 기술 도입 전략 및 기술을 연구하였다. 주요 연구 분야는 Data Mining, NLP(Natural Language Processing), STT(Speech-To-Text), TTS(Text-To-Speech), Vision 기술들의 융합 등이다.