지능형 미디어 콘텐츠 편집 기술 개발 현황

ㅁ **추연승, 김현식** / 한국전자기술연구원

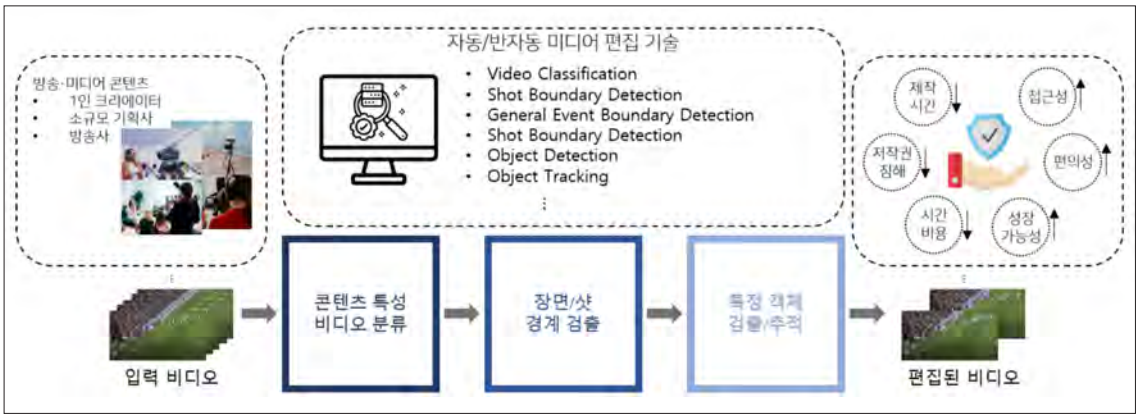
요 약

미디어 콘텐츠 편집 기술은 콘텐츠 제작 과정에 필수적으로 요구되는 기술로, 비디오 특성 기반 장르 분류 기술, 자동 장면 분할 기술, 객체 인식 기술 등으로 구분될 수 있다. 코로나19 이후 미디어 콘텐츠 시장은 폭발적으로 성장하였으며, 인공지능을 활용하여 콘텐츠를 보다 쉽게 제작하려는 요구가 증가하면서 인공지능 기반의 미디어 콘텐츠 제작 및 편집 기술에 대한 연구 개발이 활발히 진행되고 있다. 본고에서는 미디어 콘텐츠 제작 과정에 적용 가능한 인공지능 기반의 미디어 편집 기술에 대한 개발 현황에 대하여 살펴본다.

I. 서론

최근, 초고속 통신 기술의 발달로 미디어 콘텐츠에 대한 접근이 수월해짐에 따라 미디어 콘텐츠의 수요가 크게 증가하고 있다. 이에 따라 방송국 또는 광고 제작사 등 전통적인 콘텐츠 제작사 외에도 크리에이터, MCN(Multi Channel Network)과 같은 새로운 제작사가 등장하여 콘텐츠 공급 또한 급격하게 증가하고 있다. 한국의 유튜브 데이터 분석 업체 소셜러스가 발표한 '2021년 한국 유튜브 빅데이터 분석 보고서'에 따르면 2021년 유튜브 구독자 규모 및 누적 조회 수 규모는 각각 전년도 규모에 비해 148%, 176%가 증가한 36.5억 명, 1조 5103억 회로 집계됐다[1].

이렇듯 미디어 콘텐츠 수요가 눈에 띄게 증가함에 따라 공급 과정의 중요성 역시 크게 부각되고 있으며, 특히, 모든 촬영본에 대하여 검토하고 편집하는 전반적인 과정은 미디어 콘텐츠 제작자들의 주된 관심사로 급부상하였다. 일례로 촬영본을 검토하고 편집하는 일련의 과정들



<그림 1> 지능형 미디어 콘텐츠 편집 과정

은 순수하게 편집만 하는 과정을 제외하고도 주제를 찾고 편집점을 잡는 영상 분석 과정에만 원본 분량의 2배에 달하는 시간이 소모된다[2]. 그러한 부담을 최소화하기 위해, 인공지능과 컴퓨터 비전 기반의 영상 분석을 통한 지능형 영상 분석 및 편집 기술에 대한 개발이 본격적으로 이루어지고 있다. 이에 따라 본고에서는 대표적인 미디어 콘텐츠 편집 기술을 목적에 따라 크게 세 가지로 분류하여 소개한다.

첫 번째는 미디어 콘텐츠의 특성을 이해한 정보를 기반으로 장르 및 장면을 구분하기 위한 콘텐츠 특성 추출 기술이다. 이 기술은 콘텐츠로부터 추출한 정보를 통해 콘텐츠의 장르, 유형을 이해하고 분류하는 기술이다.

두 번째로, 주요 객체, 이벤트를 기준으로 편집을 하기 위해 영상 내 정보를 특정하여 장면을 구분, 분할하는 지능형 영상 가공 기술이다. 주로 장면 경계 검출(Scene Boundary Detection)과 이벤트 경계 검출(Generic Event Boundary Detection)이 지능형 영상 가공 기술에 포함된다.

마지막은, 영상 내 객체를 분석하는 객체 인식과 관련된 기술이다. 예를 들어 현재도 다수의 미디어 콘텐츠에서는 영상을 촬영하는 과정에서 행인의 얼굴이나 주행 차량의 번호판과 같은 개인정보가 노출되는 식별 정보 유출 문제가 의도치 않게 발생하고 있다. 혹은, 특정 객체를 기준으로 영상을 편집해야 하는 경우가 있기도 하다. 이와 같은 상황들은 특정 객체를 기준으로 장면 편집이 요구되므로 영상 내 객체에 대한 인식 기술이 필요하다. 미디어 콘텐츠 내 특정 객체를 인식하는 기술은 객체 검출(Object Detection) 기술과 객체 추적(Object Tracking) 기술 등이 포함된다.

본고는 다음과 같이 구성된다. 2장에서는 지정된 기준에 따라 콘텐츠를 분류하는 기술인 콘텐츠 특성 추출 기술에 대하여 소개하고, 3장에서는 분류된 콘텐츠에 대해 장면, 이벤트 단위로 경계를 검출하는 지능형 영상 가공 기술에 대해 알아본다. 4장을 통해서는 지능형 객체 인식 기술에 대해 설명하고, 마지막 5장에서는 망라한 기술들을 포함, 미디어 편집 기술에 대한 결론을 맺는다.

II. 콘텐츠 특성 추출 기술

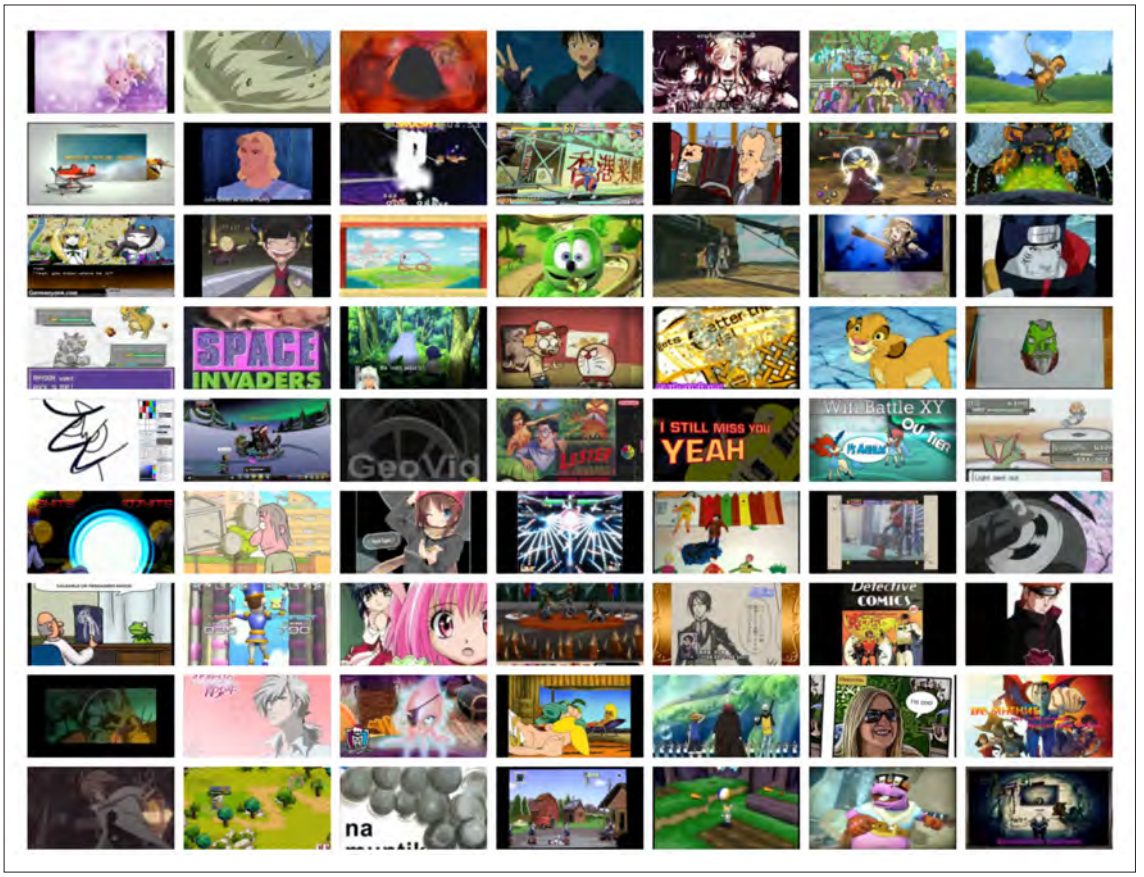
콘텐츠 특성 추출 기술은 미디어 콘텐츠가 각 장르별로 갖춘 시각적 특징을 추출하여 콘텐츠 특성을 분석하는 기술이다. 예를 들어 방송 미디어 콘텐츠는 예능, 다큐, 토론회, 스포츠와 같이 여러 장르가 있으며, 장르마다 특징이 상이하다. 각각의 비디오 속 고유의 특징과 특성, 속성들을 추출하여 분류할 수 있다면, 분류된 결과에 맞는 편집

툴이나 편집 디자인을 추천할 수 있다.

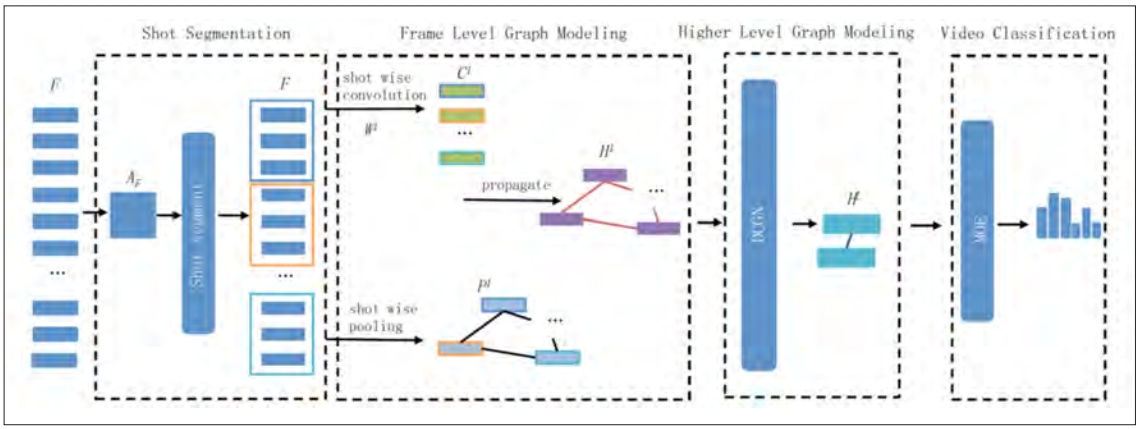
콘텐츠 특성 기반 대분류를 위한 비디오 분류 기술은 비디오 그 자체의 속성을 이해하고자 하는 비디오 분류(Video Classification)와 비디오 내 행위를 기반으로 분류를 수행하는 행동 분류(Action Classification)로 크게 두 가지 형태로 나눌 수 있다.

먼저, 비디오 분류를 위한 데이터셋과 방법은 다음과 같다. 2016년 Google Research에서는 기존에 없던 50만 시간, 8백만 개의 방대한 비디오 데이터셋 'Youtube-8M'을 배포하였다[3]. 〈그림 2〉는 Youtube-8M 데이터셋의 예시를 나타낸다. Youtube-8M은 3,800개의 주제를 갖는 비디오 동영상의 프레임에 대해 따로 계산된 오디오와 영상 특징 정보를 제공한다. Youtube-8M은 시청각적 특징 정보를 활용해 비디오 이해(Video Understanding)와 관련된 대표 데이터셋 중 하나로 활용되고 있으며, 이후 Youtube-8M 데이터셋을 활용한 다양한 방법론이 등장하였다.

2018년 F. Mao 등은 GCN(Graph Convolutional Network) 형태의 장면, 이벤트별 노드 단위 표현을 통한 비디오 분류 방법을 제안하였다[4]. 〈그림 3〉은 F. Mao 등이 제안한 방법의 전체 프레임워크를 나타낸다. 〈그림 3〉에서의 'Frame Level Graph Modeling' 부분과 같이 저자들은 대부분의 비디오 영상 내 순서가 항상 일정한 것이 아니라는 것을 근거로 기존에 주로 사용되던 LSTM(Long-Short Term Memory), GRU(Gated Recurrent Unit)와 같은 시계열 네트워크가 아닌 그래프 형태의 GCN을 활용



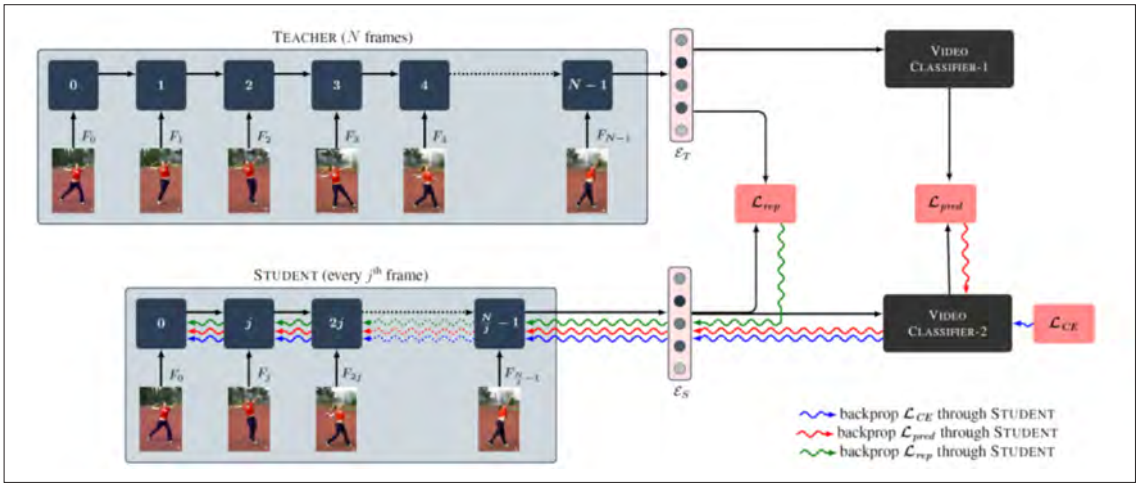
<그림 2> Youtube-8M 데이터셋의 '만화' 카테고리 예시



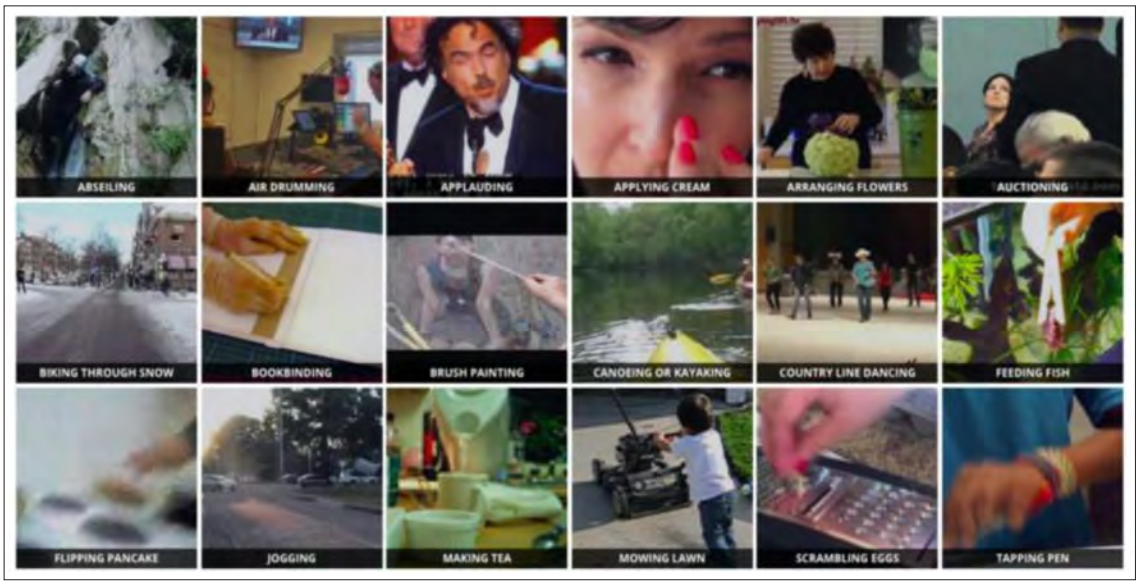
<그림 3> Hierarchical Video Frame Sequence Representation with Deep Convolutional Graph Network의 프레임워크

하였다. 이처럼 저자들은 시계열 입력에 대한 처리가 아닌, 장면 분할(Shot Segmentation)을 기준으로 장면별 유사성을 비교하여 엣지-노드(Edge-Node) 형태로 표현하고, 노드 레벨에서의 컨볼루션(Convolution) 연산 및 그래프 레벨에서의 평균 통합(Average Pooling)을 통해 최종적으로 전체 비디오 레벨의 고차원 정보까지 계층적으로 특징을 함축하는 방법에 대해 제안했다. 이 방법은 단순하게 LSTM, GRU와 같은 시계열 네트워크를 활용한 방법보다 더 좋은 성능을 나타낸다.

2019년에 S. Bhardwaj 등은 Youtube-8M 데이터셋을 활용하여, Teacher-Student 구조의 지식 이전(Knowledge Distillation) 방법을 기반으로 한 비디오 분류 방법을 〈그림 4〉와 같이 제안하였다[5]. Teacher-Student 구조의 지식 이전 기법은 미리 학습된 Teacher 네트워크에 대해 학습하고자 하는 Student 네트워크가 Teacher 네트워크와 유사한 결과를 나타내도록 학습하는 방법을 일컫는다. 저자들은 시퀀스의 모든 프레임을 입력한 Teacher 네트워크의 결과와 시퀀스의 적은 특정 프레임들을 샘플링하여 입력한 Student 네트워크의 결과를 비교하여 유사한 특징을 추출하도록 학습을 진행했다. 이 방



<그림 4> Efficient Video Classification Using Fewer Frames의 프레임워크



<그림 5> Kinetics dataset 예시

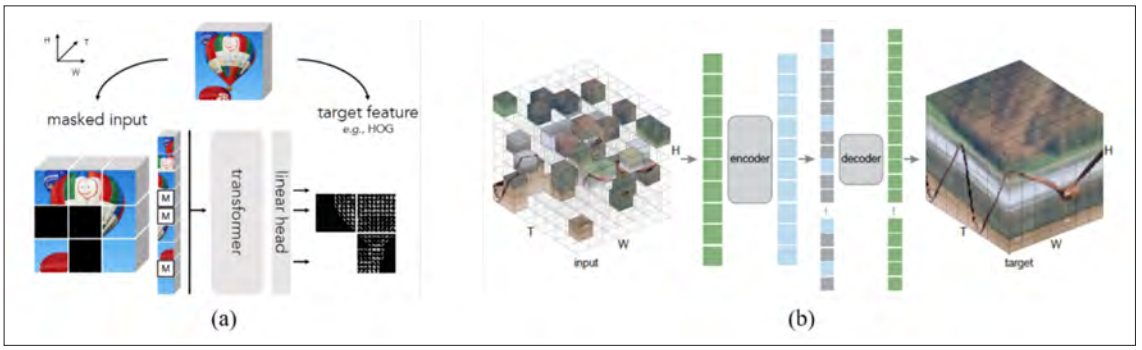
법을 통해 적은 프레임을 활용하면서도 기존의 모든 프레임을 활용한 방법과 유사한 성능을 나타낸다.

앞서 비디오 분류와 관련된 연구가 진행된 것처럼, 행동 기준의 비디오 분류도 지속적으로 연구되어 왔다. 먼저 데이터셋으로는 Kinetics, HMDB-51, UFC 101이 주로 활용된다[6-10]. 기존의 HMDB-51, UFC 101 데이터셋이 각각 51개, 101개의 행동 종류만을 정의하여 다양성이 낮았던 것에 비해, Kinetics는 400개부터 700개까지의 행동 클래스에 대해 65만 개 이상의 비디오를 제공하며 기존의 데이터셋을 대체했다. Kinetics 데이터셋은 이미지 분류 데이터셋인 이미지넷(ImageNet)처럼 비디오 이해 분야에서 사전 학습 데이터셋으로 활용되고 있다[9]. 〈그림 5〉는 Kinetics 데이터셋의 예시를 나타낸다.

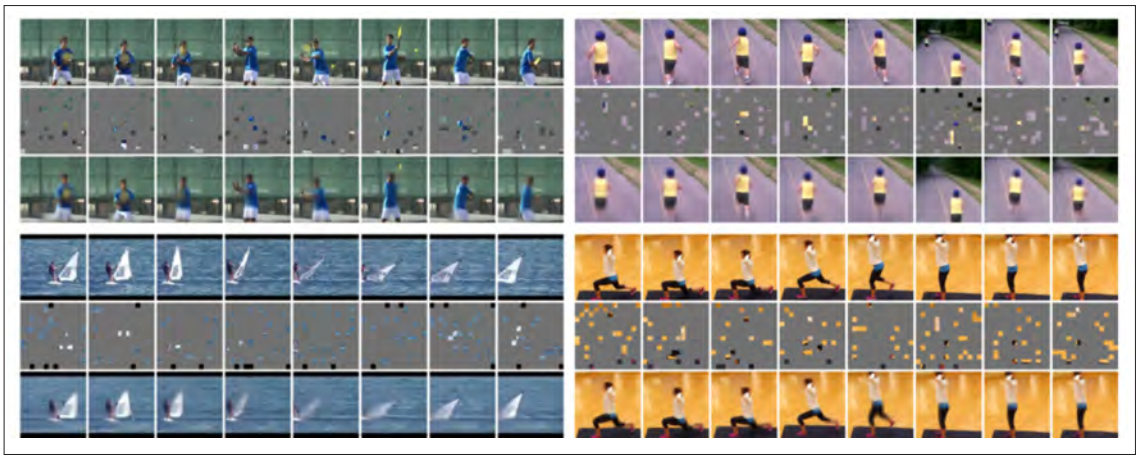
Kinetic 데이터셋을 기반으로 하는 행동 분류 기술로는 트랜스포머(Transformer) 기반의 사전 학습 방법들이 연구되고 있으며, 현재 주류로 자리 잡아 가는 추세이다. 트랜스포머 모델은 기존의 NLP(Natural Language Processing)에서 유래되어 높은 성능을 나타낸 모델로, 컴퓨터 비전 분야에서도 ViT(Vision Transformer)를 선두로 영상에서 활용 가능하도록 변형, 적용되며 근래 가장 주목받고 있는 모델 중 하나이다[33].

트랜스포머 모델을 기반으로 한 최신 활용 방법에 대해 간단히 언급하자면 다음과 같다. 먼저 2022년 C. Wei 등은 기존 입력에서 영상의 일부를 마스킹(Masking) 하고, 가장 많이 사용되는 Handcrafted-feature 중 하나인 HOG(Histogram of Oriented Gradients) 특징을 예측하는 간단한 방법을 제안했다[10-11]. 〈그림 6〉의 (a)는 C. Wei가 제안한 방법을 나타낸다. 해당 논문에서는 딥러닝 이전에 자주 쓰였던 Handcrafted-feature를 재평가하고 특히 그중에서도 HOG의 활용성을 높게 평가하며 실제로 트랜스포머를 활용한 자기 지도 학습(Self-Supervised Learning) 형태의 선행 학습 단계에서 특징 결과를 HOG와 유사하게 예측하도록 학습하는 것이 좋은 성능을 나타내는 것을 실험을 통해 확인하였다. 실제 실험 결과 Kinetics 데이터셋에서 C. Wei가 제안한 방법은 다른 방법에 비해 좋은 정확도를 기록했다.

또한 C. Feichtenhofer 등은 K. He가 제안한 MAE(Masked Auto Encoder)를 시퀀스 환경에 그대로



<그림 6> 트랜스포머를 활용한 비디오 내 행동 분류 방법. (a): C. Wei 등이 제안한 방법[11], (b): C. Feichtenhofer 등이 제안한 방법[13]



<그림 7> C. Feichtenhofer 등이 제안한 방법[13] 예시

적용한 학습 방법에 대해 제안했다[12-13]. 〈그림 6〉의 (b)는 C. Feichtenhofer 등이 제안한 방법을 나타낸다. 저자들은 〈그림 7〉과 같은 시퀀스 내 MAE 적용 실험을 통해 최소한의 도메인 관련 지식이나 귀납적 편향(inductive bias) 정보만으로도 강력한 표현 학습(Representation Learning)이 가능하다는 것과, 기존 MAE에서 강조했던 마스킹 비율의 중요성을 다시 한번 확인하였다.

이처럼 비디오 내 행동 분류 기술들은 과거에 이미지에서 딥러닝이 적용된 이미지넷 기반의 이미지 분류(Image Classification) 기술이 크게 성행했던 것처럼, 트랜스포머가 적용된 Kinetics 데이터셋 기반의 행동 분류를 주제로 선행 연구가 진행되는 것을 알 수 있다.

지금까지 미디어 콘텐츠 대분류를 위한 비디오 분류 방법들에 대해 알아보았다. 앞서 언급한 것처럼 비디오 속성에 대한 종합적인 이해를 하기 위해 대규모 데이터셋이 제작되었고, 이를 활용한 다양한 방법들이 제안됐다. 그리고 최근에는 비디오 도메인에서 트랜스포머를 활용한 선행 학습 관련 연구가 행동 분류라는 주제로 진행되고 있음을 알 수 있다.

III. 지능형 영상 가공 기술

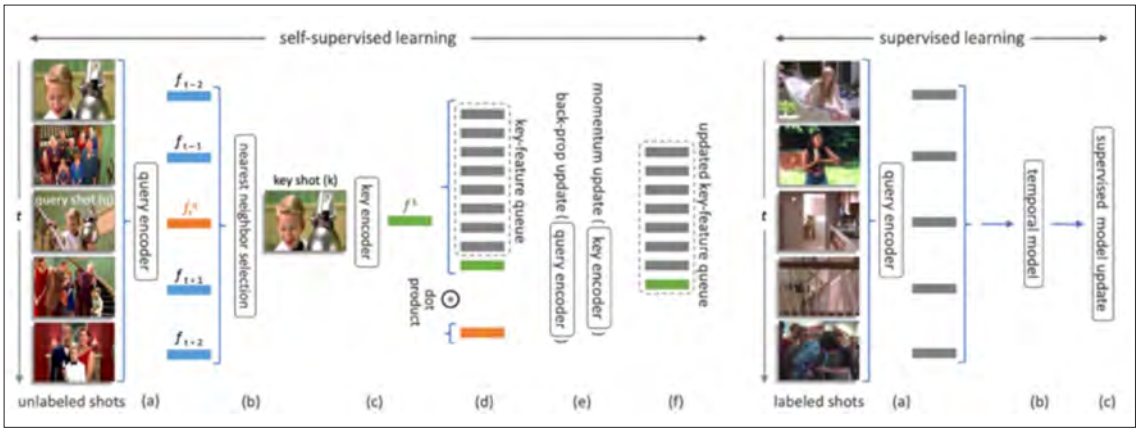
지능형 영상 가공 기술은 비디오의 흐름을 기반으로 중

요한 프레임을 추출하여 요약하거나, 구간을 지정하는 용도로 활용되는 기술을 일컫는다. 예를 들어, 장면 경계 검출(Scene Boundary Detection)은 비디오를 장면 단위로 분할하는 기술을 의미한다. 이러한 장면 분할 기술은 긴 영상 정보의 특징 집합을 분류하여 장면으로 나눠야 하기 때문에 비디오의 흐름에 대한 이해를 바탕으로 하는 고난도의 기술이 요구된다. 반면, 전체 비디오에 대해 단순한 이벤트(카메라 각도, 조도, 객체 변화 등)를 기준으로 경계를 나누는 이벤트 경계 검출(Generic Event Boundary Detection)도 있다. 따라서 본 장에서는 장면 경계 검출과 이벤트 경계 검출 두 기술에 대해 알아보고자 한다.

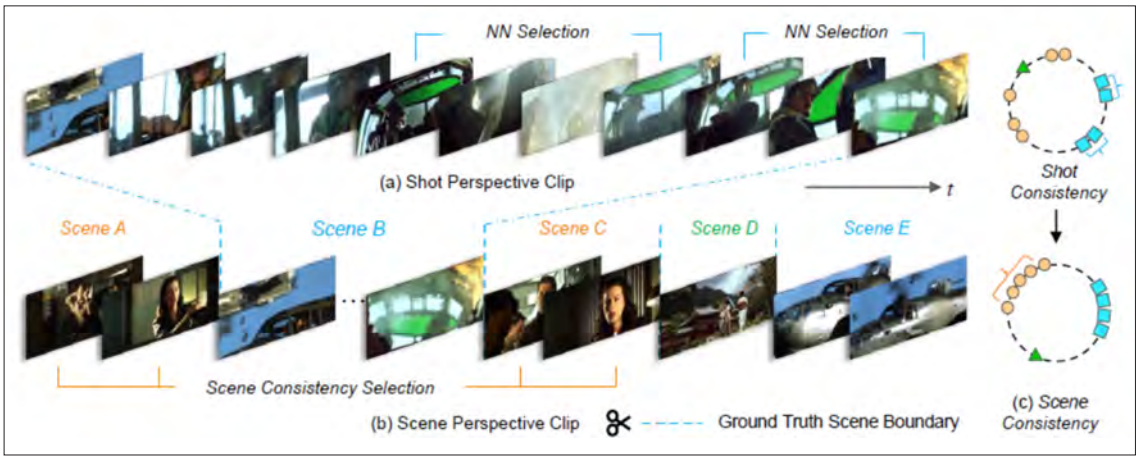
비디오의 장면 경계 검출은 MovieNet 데이터셋 등이 사용된다[14]. MovieNet은 1,100여 개의 영화에 대한 데이터셋으로, 〈그림 8〉과 같이 4만 2천여 개의 장면 경계를 포함하여 6만 5천여 개의 장소 및 장면에 대한 태그와 9만 2천여 개의 영화 스타일에 대한 태그 등의 정보를 담고 있다.

Amazon의 S. Chen 등은 2021년 MovieNet 데이터셋을 사용해 MoCo를 차용한 장면 경계 검출 방법론에 대한 논문을 공개하였다[15-16]. MoCo는 2020년 CVPR에서 K. He가 처음 제안한 대조 학습(Contrastive Learning)의 한 방법으로, 기존의 사전 형태의 학습 방법을 유연하게 사용하기 위해 사전 내 데이터의 업데이트를 효율적이게 할 수 있도록 제안한 방법이다. MoCo는 'Momentum Encoder'라는 이 개념을 통해 적은 배치 사이즈(Batch Size)로도 충분한 양의 부정 데이터셋(Negative Sample)을 활용한 학습을 할 수 있었다. S. Chen 등 저자들은 MoCo의 방식을 그대로 장면 단위에 적용하여 장면 간 유

<그림 8> MovieNet 데이터셋 예시



<그림 9> "Shot Contrastive Self-Supervised Learning for Scene Boundary Detection" 프레임워크



<그림 10> "Scene Consistency Representation Learning for Video Scene Segmentation" 프레임워크

사성을 추정할 수 있는 자기 지도 학습(Self-Supervised Learning) 과정과 장면 경계를 판단하는 MLP(Multi-Layer Perceptron)를 학습하는 지도 학습(Supervised Learning) 과정을 혼합한 장면 경계 검출 방법을 제안했고, 이는 기존의 방법들에 비해 좋은 성능을 나타냈다. 〈그림 9〉는 저자들이 제안한 자기 지도 학습과 지도 학습 과정을 나타낸다.

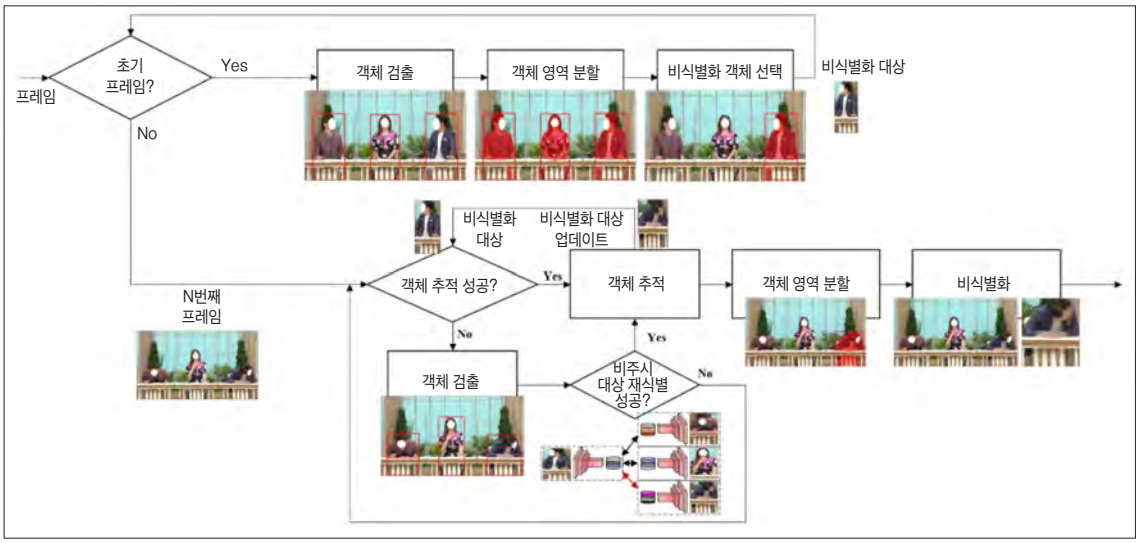
다음으로 H. Wu 등은 대조 학습 과정에서 같은 장면 내에서도 유사하지 않은 샷의 존재로 인해 장면의 연속성이 소실됨에 따라 학습이 잘 수행되지 않는 문제점을 지적하며, 장면 내부만이 아니라 다른 장면의 샷을 검색하여 유사한 긍정 데이터셋(Positive Sample)을 가져오는 전략을 통해 장면 특징 추출 과정에서 일관성을 유지하는 방법을 〈그림 10〉과 같이 제안했다[17]. 또한, 이렇게 장면 구분이 아닌 샷의 집합으로 학습 과정에서 장면의 경계로 인해 발생하는 귀납적 편향에 상대적으로 자유롭다는 이점을 취할 수 있음을 실험을 통해 증명하였다.

이 외에도 이벤트 경계 검출과 관련된 방법이 있다. 위에서 언급한 바와 같이, 장면 경계 검출은 긴 시퀀스 정보의 스토리 라인을 이해하고 장면별로 나눈 것이라면, 이벤트 경계 검출 기술은 비디오 내 모든 이벤트 변화를 감지해 경계를 구분해 내는 과제로, 2021년 M. Z. Shou 등에 의해 새롭게 정의된 분야이다[18]. M. Z. Shou 등이 제안한 Kinetics-GEBD 데이터셋을 통해 주로 연구가 진행되고 있는 이벤트 경계 검출 기술은 기존의 장면 경계 검출 분야와 마찬가지로 표현 학습, 그중에서도 대조 학습을 활용한 방법들이 연구가 진행되고 있다[19-20].

우리는 앞서 장면, 이벤트 단위의 경계 검출에 대한 다양한 방법들이 제안된 것을 확인하였다. 물론 현재에도 비디오 요약, 비디오 클립 생성, 썸네일 추출과 같은 유사 서비스가 제공되고 있는 것은 사실이다. 그러나 아직까지 거대한 비디오 데이터에 대해선 편집자가 일일이 경계 추출 과정을 수행하는 현실적인 문제점이 존재하는 만큼, 자동 편집을 위한 기술 개발이 꾸준히 요구될 것으로 보인다.

IV. 객체 인식 기술

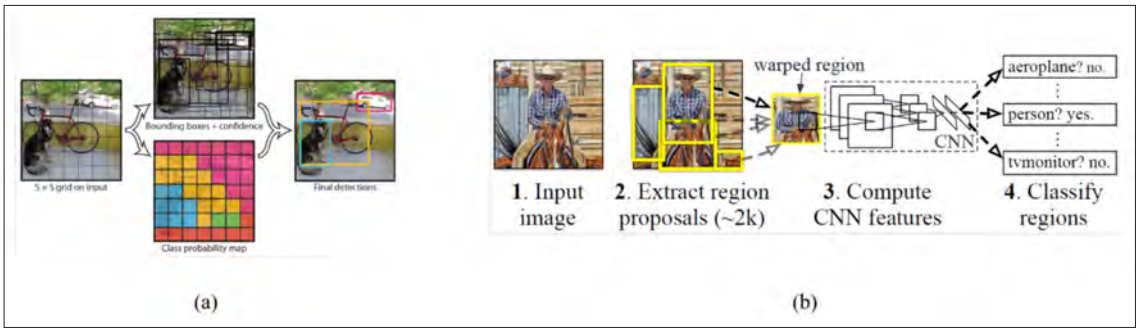
콘텐츠 특성 추출 기술을 통해 비디오 특성을 이해, 분류하고 지능형 영상 가공 기술을 통해 콘텐츠를 더욱 세분화된 장면, 이벤트, 샷 단위로 분류하면, 분류된 영상에 대해 각 특성에 맞는 객체 인식 과정을 적용할 수 있다. 객체 인식이 필요한 예로는 객체 비식별화 과정이 있다. 미



<그림 11> 장면 내 비주시 대상에 대한 비식별화 자동화 과정 예시

디어 콘텐츠에서는 당연하게도 의도치 않는 개인 식별 정보가 나타나는 경우가 있는데 이에 대해 편집자가 시퀀스 내 객체를 찾아 수동 비식별화를 진행하는 것은 굉장히 비효율적이다. 따라서, 〈그림 11〉과 같이 일련의 비식별화 자동화 과정이 필요하다. 그러나 이처럼 객체와 관련된 모든 편집 과정은 객체 인식 결과를 기반으로 진행되기 때문에, 객체 인식 기술은 미디어 편집 과정에서 중요한 기술 중 하나이다. 따라서 본 장에서는 객체 기준의 미디어 편집을 위해 필수 단계인 객체 검출 및 추적 기술에 대해 알아보고자 한다.

먼저 객체 검출은 컴퓨터 비전 내 가장 활발한 분야 중 하나이다. PASCAL VOC challenge, ILSVRC(ImageNet Large Scale Visual Recognition Challenge)로 발현한 딥러닝 이슈는 AlexNet, VGGNet, ResNet 등을 거치며 이미지 분류로서 단기간에 빠르게 발전하였다[21-25]. 이러한 영향은 본격적으로 이미지 분류와 유사한 다른 분야로 파생되기 시작했고, 이미지 분류에 국소적 정보(Localization)가 추가된 객체 검출 기술도 그중 하나이다. 이후 객체 검출 기술은 크게 2가지 형태로 진행이 되었는데, 먼저 하나는 YOLO(You Only Look Once)와 같이 경계 영역과 이미지의 클래스 분류 확률을 동시에 예측해 처리하는 단일 스테이지 형태의 검출기다(〈그림 12〉의



<그림 12> (a): YOLO 객체 검출 과정, (b): 2단계로 이루어진 R-CNN의 객체 검출 과정

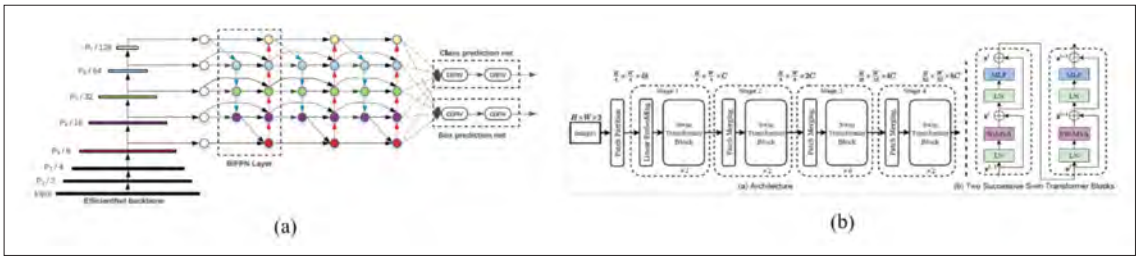
(a)). 단일 스테이지 객체 검출기는 객체 검출 시간을 크게 단축시킬 수 있으나, 정확도가 2스테이지의 검출기에 비해 낮다는 단점이 있다. 반대로 관심 영역 추정 이후 객체 분류를 수행하는 2스테이지 형태의 R-CNN(〈그림 12〉의 (b)) 시리즈는 연산 시간은 오래 걸리나, 상대적으로 정확하다는 특징이 있다[26-29].

최근에는 〈그림 13〉의 (a)와 같이 검출 모델의 효용성을 극대화하거나, 〈그림 13〉의 (b)처럼 트랜스포머를 활용해 어텐션 기법을 적용하려는 연구들이 주를 이루고 있다. M. Tan 등은 CNN(Convolutional Neural Network)의 깊이, 채널 수, 해상도와 같은 근본적인 요소에 대한 고찰로 크게 주목받은 EfficientNet을 활용한 객체 검출 기법인 EfficientDet을 발표했다[30-31]. 논문에서 M. Tan 등은 기존 스케일 변화에 강건한 특징 추출을 위한 피라미드 구조의 네트워크 FPN(Feature Pyramid Network)을 개선한 BiFPN(Bi-directional Feature Pyramid Network)이라는 개념을 통해 스케일 변화 측면에서 기존의 방법보다 강건한 검출 성능을 나타냈다. 또한 EfficientNet에서 사용된 깊이, 너비, 해상도를 적당히 조절하며 최상의 성능을 추출하려 했던 Compound Scaling 과정을 EfficientDet에서 고유의 형태로 바꿔 적용하였다. 그 결과, EfficientDet은 더 작은 모델, 더 적은 연산으로도 다른 모델에 비해 COCO 데이터셋에서 좋은 결과를 나타냈다[32].

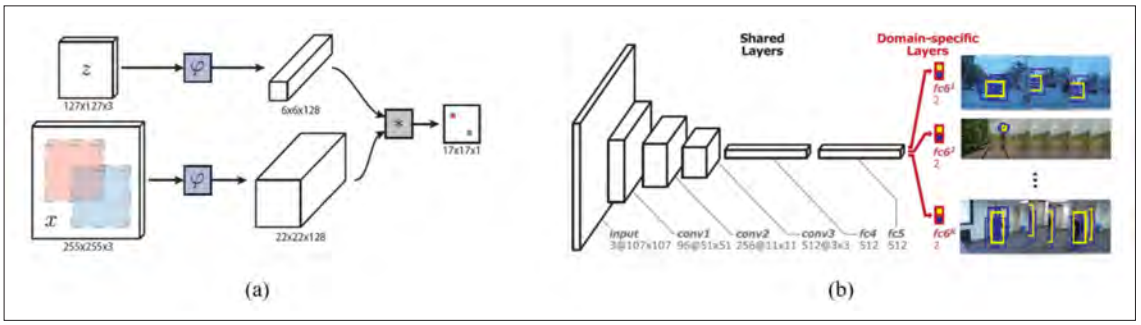
EfficientDet이 객체 검출을 위한 최적의 모델 구현을 목적으로 진행된 연구라면, 트랜스포머와 같은 대용량 파라미터를 활용해 절대적인 객체 검출 성능을 기록하기 위한 목적으로 진행된 연구도 있다. Swin Transformer는 Z. Liu 등이 제안한 방법으로 기존의 ViT의 방식과 차별화를 둔 점이 특징이다[33-34]. ViT는 이미지를 패치 단위로 분할하여 셀프 어텐션(self-attention) 기법을 적용한 단순한 방법이 사용된 반면에, Swin Transformer는 스케일 피라미드(Scale Pyramid)나 FPN과 유사한 형태로 접근하여 기존 ViT 단점인 크기, 해상도 변화에 따른 성능 저하 문제를 해결하고자 하였다. 이를 위해 우선 Swin Transformer는 작은 크기의 패치로 시작해서 점점 패치를 병합해 크기를 키워나가는 계층적 구조의 패치 활용 방법으로 스케일 문제를 해결하려 하였다. 또한 패치의 집합인 윈도우 단위의 셀프 어텐션 기법이 단순 적용된 기존 방법에 비해, Swin Transformer는 고정된 윈도우는 윈도우의 경계 부분에서는 어텐션 기법이 소외되는 문제를 해결하기 위해 윈도우를 바꿔가며 어텐션을 적용하는 Shifted-Window 개념을 제시했다. 이 두 핵심 방법을 통해 Swin Transformer는 객체 검출 분야에서 좋은 성능을 기록했다.

다음으로는 검출된 객체에 대한 추적 과정이 필요하다. 객체 검출과 마찬가지로, 객체 추적 기법도 딥러닝 이후 변화된 것들이 많다. 전통적인 객체 추적 방법은 옵티컬 플로우(Optical Flow)와 같은 픽셀 레벨의 변화를 검출하거나, 칼만 필터(Kalman Filter), 파티클 필터(Particle Filter)처럼 상태를 예측하여 객체를 추적하는 방법이 활용됐다. 그러나 딥러닝이 본격적으로 파생된 이후, 딥러닝 아키텍처를 활용한 객체 추적 방법들이 제안되었다.

먼저 샴 네트워크(Siamese Network) 기반의 방법이

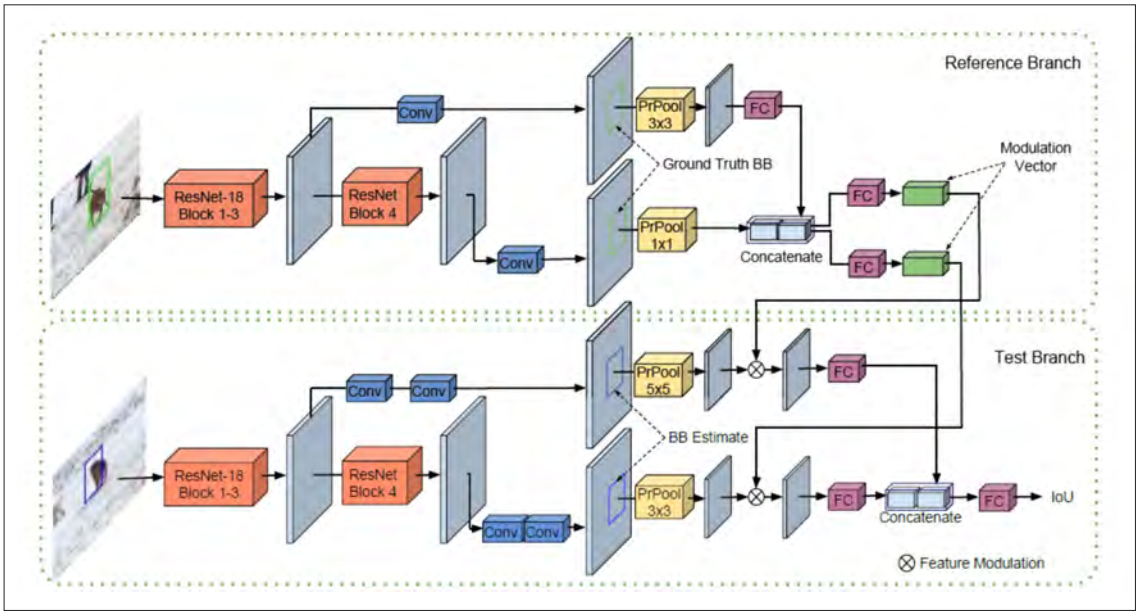


<그림 13> (a): EfficientDet의 구조, (b): Swin Transformer의 구조



<그림 14> (a): SiameseFC 구조, (b): MDNet 구조

있다. L. Bertinetto 등은 Yann Lecun 교수팀이 2005년 제창한 샴 네트워크를 객체 추적 과정에 활용하는 SiameseFC를 제안했다[35-36]. 샴 네트워크는 말 그대로 동일한 네트워크를 활용하는 방법으로, 동일한 네트워크에 다른 입력을 넣어서 나오는 특징 차이를 활용하기 위한 개념이다. L. Bertinetto 등은 시간대가 다른 영상을 입력으로 넣었을 때의 결과에서 유사성을 비교하여 객체 추적에서 좋은 결과를 나타낼 수 있었다. 이렇게 샴 네트워크를 사용한 방법이 좋은 성능을 거두게 됨에 따라서, 샴 네트워크 기반의 다양한 객체 추적 기법들이 제안되었다[37-38]. 비슷한 시기에 H. Nam 등은 CNN의 마지막 FCL(fully connected layer)을 제외한 공유 네트워크를 활용해 다중 도메인 학습(multi-domain learning)을 적용, 공통 특징을 학습하는 방법인 MDNet을 제안하였다[39]. MDNet은 CNN과 online-offline 학습이 결합되어 성공적으로 객체 추적에 적용된 사례로, 객체 추적 분야에서 새로운 가이드라인이 되었다. 〈그림 14〉는 SiameseFC 및 MDNet 구조를 나타낸다.



<그림 15> ATOM의 구조

이후 대표되는 객체 추적 기술은 2019년 M. Danelljan 등이 CVPR에서 발표한 ATOM이 있다[40]. 해당 논문에서 저자들은 추적기의 추적하고자 하는 대상 객체에 대한 깊은 지식과 이해 없이 추적 정확도만 향상시키려는 시도를 근본적인 문제점으로 지적하며, 이 부분을 해결하기 위한 연구를 진행했다. 그리하여 ATOM은 기존 영상과 대상 영상에서 추정한 각 영역 상자(Bounding box)의 겹침 영역이 최대가 되도록 학습하는 방법을 제안했다. 〈그림 15〉는 ATOM의 구조를 나타낸다. 이 방법을 통해 ATOM은 단순히 객체를 잘 추적하려는 시도가 아니라 대상의 적절한 영역 상자가 무엇인지에 대해 고려하기 때문에, 추적 성능이 향상될 수 있다고 주장했다. 실제로 ATOM은 이 방법을 통해 VOT2018 데이터셋을 포함한 여러 데이터셋에서 좋은 성능을 기록했다[41]. 이후 TransTrack, Transformer Tracking과 같은 방법들이 등장하며 다른 비전 분야와 마찬가지로 트랜스포머 구조가 적용된 기술에 대한 연구가 지속적으로 진행되고 있다[42-43].

지금까지 객체 검출 기술과 추적 기술에 대해 간단하게 알아보았다. 기존의 Handcrafted-feature를 사용하던 객체 검출, 추적 기술은 딥러닝 이후 큰 변화를 맞이하였으며, 현재는 목표에 따라 실시간성 및 최적화 위주의 연구 혹은 절대적 성능 향상 위주의 연구로 세분화되어 가고 있는 추세이다. 따라서 객체 인식 기술은 다양한 방법들이 존재하는 만큼, 필요에 따라 적절한 방법을 사용하는 것이 중요하다.

V. 결론

지금까지 미디어 콘텐츠 제작에 많은 시간과 비용이 소요되고 있는 촬영-편집 단계에서 콘텐츠 제작에 다양하게 활용할 수 있는 정보 자원 제공 및 미디어 제작 시간을 줄여주고 제작·편집이 용이하도록 연구된 인공지능 기반의 미디어 제작·편집 기술들에 대해 알아보았다.

미디어 콘텐츠 시장의 폭발적인 성장과 디지털 미디어 콘텐츠 수요 확산에 능동적으로 대응하기 위해서 지능형 미디어 콘텐츠 편집 기술은 선택이 아닌 필수적인 기술이다. 그러나 인공지능이 미디어 콘텐츠의 기획·제작 의도에 따라 영상을 해석하고 의도가 반영되도록 자동으로 제작·편집을 하기까지는 많이 부족한 것이 현실이므로 미디어 산업의 활성화와 기술 경쟁력을 확보할 수 있도록 지능형 미디어 콘텐츠 편집 기술을 포함한 지능형 미디어 크리에이터 기술에 대한 더 많은 연구와 노력을 기울여야 할 것이다.

※ 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00804, 학습 기반 연출 기법이 적용된 미디어 제작 기술 개발)

참 고 문 헌

[1] 한국 최초/최대 유튜브 채널분석 소셜러스, "2021 한국 유튜브 분석 보고서," https://socialerus.com/
[2] 최은서, "20시간 일하고 3만 원... 계약서 한 장 없이 헐값 시장에 방치된 영상 편집자들," 한국일보, 2022년 2월 5일자.
[3] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A Large-Scale Video Classification Benchmark," *arXiv preprint arXiv:1609.08675*, 2016.
[4] F. Mao, X. Wu, H. Xue, and R. Zhang, "Hierarchical Video Frame Sequence Representation with Deep Convolutional Graph Network," in *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*, 2018.
[5] S. Bhardwaj, M. Srinivasan, M. M. Khapra, "Efficient Video Classification using Fewer Frames," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[6] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[7] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A Short Note about Kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.
[8] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A Short Note on the Kinetics-700 Human Action Dataset," *arXiv preprint arXiv:1907.06987*, 2019.
[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
[10] N. Dalal and B. Triggs, "Histogram of Oriented Gradients for Human Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
[11] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked Feature Prediction for Self-Supervised Visual Pre-Training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and Ross Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[13] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked Autoencoders As Spatiotemporal Learners," *arXiv preprint arXiv:2205.09113*, 2022.
[14] Q. Huang, Y. xiong, A. Rao, J. Wang, and D. Lin, "MovieNet: A Holistic Dataset for Movie Understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[16] S. Chen, X. Nie, D. Fan, D. Zhang, V. Bhat, and R. Hamid, "Shot Contrastive Self-Supervised Learning for Scene Boundary Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[17] H. Wu, K. Chen, Y. Luo, R. Qiao, B. Ren, H. Liu, W. Xie, and L. Shen, "Scene Consistency Representation Learning for Video Scene Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[18] M. Z. Shou, S. W. Lei, W. Wang, D. Ghadiyaram, and M. Feiszli, "Generic Event Boundary Detection: A Benchmark for Event Segmentation," in *Proceedings of the IEEE International Conference on Computer (ICCV)*, 2021.
[19] C. Li, X. Wang, L. Wen, D. Hong, T. Luo, and L. Zhang, "End-to-End Compressed Video Representation Learning for Generic Event Boundary Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[20] H. Kang, J. Kim, T. Kim, and S. J. Kim, "UBoCo: Unsupervised Boundary Contrastive Learning for Generic Event Boundary Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[21] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal Visual Object Class (VOC) Challenge," in *International Journal of Computer Vision (IJCV)*, 2010.
[22] O. Russakovsky, J. Deng, H. su, J. Krause, S. Satheesh, S. ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," in *International Journal of Computer Vision (IJCV)*, 2015.
[23] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

참 고 문 헌

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[27] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer (ICCV)*, 2015.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] M. Tan, and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[31] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proceedings of the IEEE International Conference on Computer (ICCV)*, 2021.

[35] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-Convolutional Siamese Networks for Object Tracking," in *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*, 2016.

[36] S. Chopra, R. Hadsell, and Y. Lecun, "Learning A Similarity Metric Discriminatively, with Application to Face Verification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[37] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware Siamese Networks for Visual Object Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[39] H. Nam, and B. Han, "Learning Multi-Domain Convolutional Neural Networks for Visual Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[40] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate Tracking by Overlap Maximization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] M. Kristan, and et al, "The sixth Visual Object Tracking VOT2018 challenge results," in *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*, 2018.

[42] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "TransTrack: Multiple Object Tracking with Transformer," in *arXiv preprint arXiv:2012.15460*, 2020.

[43] X. chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

저 자 소 개

추 연 승

- 2020년 : 중앙대학교 영상학과 석사
- 2020년 ~ 현재 : 한국전자기술연구원 연구원
- 주관심분야 : 영상처리, 컴퓨터비전, 인공지능

김 현 식

- 2017년 : 연세대학교 전자전기공학 박사
- 2004년 ~ 현재 : 한국전자기술연구원 수석연구원
- 주관심분야 : 블록체인, 실감미디어, 인공지능