온라인 행동 탐지 기술 동향

Trends in Online Action Detection in Streaming Videos

문진영 (J.Y. Moon, jymoon@etri.re.kr) 시각지능연구실 책임연구원
김형일 (H.I. Kim, hikim@etri.re.kr) 시각지능연구실 선임연구원
이용주 (Y.J. Lee, yongju@etri.re.kr) 시각지능연구실 책임연구원/실장

ABSTRACT

Online action detection (OAD) in a streaming video is an attractive research area that has aroused interest lately. Although most studies for action understanding have considered action recognition in well-trimmed videos and offline temporal action detection in untrimmed videos, online action detection methods are required to monitor action occurrences in streaming videos. OAD predicts action probabilities for a current frame or frame sequence using a fixed-sized video segment, including past and current frames. In this article, we discuss deep learning-based OAD models. In addition, we investigated OAD evaluation methodologies, including benchmark datasets and performance measures, and compared the performances of the presented OAD models.

KEYWORDS 온라인 행동 탐지, 비디오 행동 탐지, 비디오 행동 이해

Ⅰ. 서론

비디오 행동 이해 기술은 크게 비디오 내에 하나의 행동 인스턴스(Action Instance)만을 포함하도록 잘 분할 편집한 비디오(Well-trimmed Videos)에 대해서 행동 클래스(Action Class)를 분류하는 행동 인식(Action Recognition) 기술과 행동과 관련 없는 백그라운드(Background) 부분들과 다수의 행동 클래스에 속하는 복수 개의 행동 인스턴스를 포함하는 무편집 비디오(Untrimmed Videos)에 대해서 각 행동 인스턴스별 발생 구간의 위치를 추정하고, 행동 클래스를 분류하는 행동 탐지(Action Detection) 기술로 구분된다(그림 1 참조). 행동 탐지 중에서 행동 인식을 제외하고, 특정 행동 클래스에 상관없이 행동 인스턴스의 위치를 추정하는 기술을 행동 국지화(Action Localization) 또는 행동 프로포절 생성 기술(Action Proposal Generation)이라고 한다. 기존의 행동 탐지 방법들 중에서 일부는 행동 인식과 행동

* DOI: https://doi.org/10.22648/ETRI.2021.J.360208

* 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No.2020-0-00004, 장기 시각 메모리 네트워크 기반의 예지형 시각지능 핵심기술 개발, No.B0101-15-0266, 실시간 대규모 영상 데이터 이해·예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발].

그림 1 행동 인식 및 탐지 기술 개념도

국지화 기술을 종단 간 학습 기법으로 행동 프로포절 생성과 행동 인식을 동시에 최적화시키도록 탐지 모델을 학습시키고, 일부는 행동 프로포절 생성 기술과 프로포절에 대한 행동 인식 기술을 별도의 모델로 나누어 학습시키기도 한다. 무편집 비디오에서 하나의 행동 인스턴스를 포함하도록 비디오를 잘 분할하는 것 자체가 어려운 문제이고, 대부분의 현실 세계에서 사용자들이 소비하는 비디오들은 무편집 비디오이기 때문에 실세계 비디오에서 행동을 이해하기 위해서는 행동 탐지 또는 국지화 기술이 필수적이다.

행동 탐지 기술은 위치 추정의 대상에 따라, 위치를 시간에 국한하여 행동 인스턴스의 시간적 위치 정보를 제공하는 시간적 행동 탐지(Temporal Action Detection) 기술과 위치 추정 대상을 시·공간 모두를 고려하여 시간적 위치와 프레임 내에서의 공간적 위치 정보를 제공하는 시·공간적 행동 탐지(Spatio-temporal Action Detection) 기술로 구분된다. 그리고 행동 탐지는 입력 비디오의 처리 방식에 따라서 오프라인 행동 탐지(Offline Action Detection)와 온라인 행동 탐지(Onilne Action Detection)로 구분된다. 오프라인 행동 탐지 기술은 무편집 비디오 전체를 입력으로 주고, 입력 비디오에 포함된 복수 개의 행동 인스턴스들의 위치와 행동 클래스를 출력으로 제공하는 데 반해, 온라인 행동 탐지는 스트리밍 비디오를 등간격으로 분할한 비디오 세그먼트(Video Segment)를 입력으로 매 프레임 행동 클래스 예측 결과를 제공하는 것을 목표로 한다(그림 1).

지금까지 대부분의 비디오 행동 탐지 연구들이 오프라인 방식의 시간적 행동 탐지 분야에 집중되어 있지만, 스트리밍 비디오를 위한 온라인 행동 탐지에 대한 관심이 점점 커지고 있다. 지능형 CCTV와 같이 장시간 대용량 비디오에 대한 지속적인 모니터링을 위해서는 행동 발생 직후 통지 기능이 필요하다. 이를 위해서는 비디오 전체가 아닌 순차적으로 들어오는 스트리밍 비디오를 처리 가능해야 한다. 따라서 본 고에서는 스트리밍 비디오를 위한 온라인 방식의 시간적 행동 탐지 기술을 다룬다.

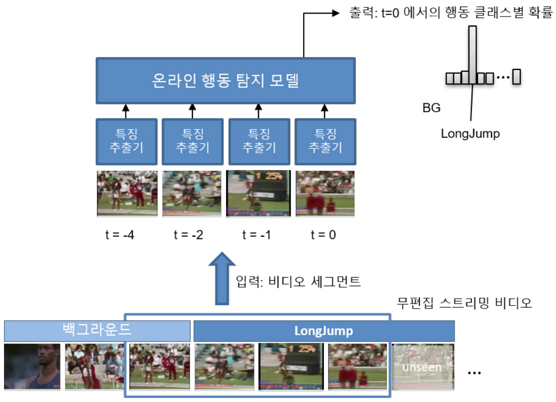

그림 2 **온라인 행동 탐지 모델 기본 구성도**

Ⅱ. 온라인 행동 탐지 기술 개요

비디오 전체를 입력 받아 행동 구간의 시작과 종료를 탐지하는 기존 오프라인 행동 탐지와 달리, 온라인 행동 탐지는 지속적으로 입력되는 스트리밍 비디오에서 과거부터 현재까지 관찰된 하나의 비디오 세그먼트에서 현재 프레임 또는 프레임 시퀀스의 행동을 탐지한다.

입력이 비디오 전체가 아니라서 비디오 세그먼트 내의 행동은 행동의 일부분만을 포함할 수 있고, 완결된 행동은 여러 개의 비디오 세그먼트에 걸쳐서 발생할 수 있기 때문에 오프라인 행동 탐지와 같이 행동의 시작과 종료를 탐지 결과로 출력할 수는 없다. 그래서 온라인 행동 탐지 모델은 현재 탐지 목표로 하는 행동들의 길이를 고려해서 정의한 길이의 비디오 세그먼트를 하나의 단위로 입력받아 행동 세그먼트 내의 정보를 이용해 최근 프레임에서의 행동 분포를 획득한다(그림 2).

Ⅲ. 온라인 행동 탐지 방법

이 장에서는 딥러닝 기반의 온라인 행동 탐지 방법들에 대하여 살펴보기로 한다.

1. RED 모델

Reinforced Encoder–Decoder(RED)[1]는 강화 학습 기법으로 입력 비디오 세그먼트의 시각적 표현(Visual Representations)을 인코딩하고, 미래 프레임들의 시각적 표현을 예측하여 현재 행동을 탐지하거나 미래 행동을 예측하는 모델이다. 이를 위해 RED[1]는 현재 프레임들의 시각적 표현을 인코딩하고 미래 프레임들의 시각적 표현을 예측하는 인코딩–디코딩 네트워크, 예측된 미래의 시각적 표현을 이용해서 미래 행동 클래스를 예측하는 분류 네트워크, 가능한 한 빨리 행동 클래스를 예측할수록 보상을 더 주는 강화 학습 모듈로 구성된다(참고문헌 [1]의 그림 2). 즉, RED[1]에서는 처리하는 비디오 세그먼트 내의 프레임별 시각 정보를 LSTM으로 축적하고, 결과를 바로 FC(Fully Connected) 레이어를 이용해 현재 행동을 탐지하는 것이 아니라 미래 프레임들의 시각 특징을 예측해서 이를 통해 현재 행동을 탐지하는 것에 초점을 맞추고 있다.

RED[1]의 학습은 크게 두 단계로 구성된다. 첫 번째 학습 단계에서 미래 프레임들의 예측된 시각적 표현들과 실제 시각적 표현들 간의 오차를 줄이는 회귀 손실함수에 의해 최적화된다. 행동 클래스가 필요 없기 때문에 레이블이 없는 비디오 세그먼트를 이용해서 인코딩–디코딩 네트워크를 충분히 학습시킬 수 있다. 두 번째 학습 단계에서는 회귀 손실함수와 더불어, 예측한 행동 클래스에 대한 분류 손실함수와 강화 모듈의 손실함수를 함께 고려하는 통합 손실함수에 의해 최적화된다.

RED[2]는 실험에서 0.25초 길이에 해당하는 연속된 6개 프레임을 하나의 단위로 시각적 표현을 생성하는데, 인코더에서는 4초를 입력으로 인코딩해서 디코더에서 2초 길이의 시각적 표현을 예측

하였다. 그래서 RED[2]는 0.25초에서 2초 사이의 행동 클래스를 예측하는데, 그 중에서 최소 예측 길이인 0.25초 뒤의 행동 클래스 예측 결과를 온라인 행동 탐지 결과로 제시하였다.

2. TRN 모델

Temporal Recurrent Network(TRN)[2]는 TRN 셀이라는 온라인 행동 탐지를 위한 RNN 네트워크를 위한 새로운 유닛을 제안하고, 이를 기반으로 하는 네트워크를 통해 입력 비디오 세그먼트에 대한 현재 행동 클래스를 예측한다.

과거 정보들을 축적하는 GRU나 LSTM과 달리, TRN 셀은 과거 증거들뿐만 아니라 예측한 미래 정보를 함께 축적시키는 역할을 한다. TRN 셀은 크게 시간적 디코더, 미래 게이트, 그리고 시·공간 축적기 3개의 컴포넌트로 구성된다. 시간적 디코더는 시각적 표현을 학습하여 미래 시퀀스의 행동을 예측한다. 미래 게이트는 디코더로부터 히든 상태의 벡터를 받아서 미래 상황정보로 이 특징들을 임베딩시킨다. 그리고 시·공간 축적기에서는 이전, 현재, 미래 정보로부터 시·공간 특징을 캡처하고, 현재 프레임에서 일어나고 있는 행동 클래스를 예측한다(참고문헌 [2]의 그림 2).

미래 정보를 예측하기 위해 TRN[2]에서는 학습 시에 현재 프레임과 미래 프레임을 모두 포함한 비디오 전체를 입력으로 사용한다. 그리고 테스트 시에는 미래 프레임에 액세스하지 않고, 예측된 미래 정보만 사용하여 온라인으로 행동을 탐지한다.

3. IDN 모델

TRN[2]과 유사하게, Information Discrimination Network(IDN)[3]는 비디오 세그먼트에서 각 프레임의 시각 정보를 입력으로 새로 제안한 RNN 유닛인 Information Discrimination Unit(IDU)를 통해 최신 현재 프레임의 행동 클래스를 예측한다. 기존 RNN 유닛들과 달리, 처리 시점과 현재 시점의 2개의 시각 특징들을 입력으로 받는다. 기존 RED[1]와 TRN[2]과 달리, IDN[3]은 한 비디오 세그먼트 내에는 최신 현재 시점의 행동과 관련된 행동과 관련되지 않은 행동들이 혼재할 수 있다는 점에 초점을 맞추어, 비디오 세그먼트 내의 모든 시각 정보를 동일한 중요도가 아닌 현재 행동과 관련된 프레임과 관련되지 않은 프레임의 시각 정보 간의 차별을 두어 행동 정보를 축적하는 방법을 제안한다.

IDU는 매 시점의 정보를 축적시키는 GRU에 파란색 선에 해당되는 부분과 붉은색 상자로 표시된 얼리 임베딩(Early Embedding) 모듈이 추가되어 있다(참고문헌 [3]의 그림 2). 리셋 모듈과 업데이트 모듈은 과거 정보와 현재 행동 간의 관계를 모델링하는 역할을 한다. 리셋 모듈은 현재 행동에 따라서 효과적으로 과거 정보를 떨어뜨리거나 취하여 다음 단계로 정보를 넘기도록 해 준다. 업데이트 모듈은 시점의 시각 정보와 최근 현재 시점의 시각 정보 간의 관련성을 고려해서 업데이트를 처리한다. 얼리 임베딩 모듈은 입력되는 특징이 행동 인식 네트워크에서 추출되어 행동 탐지보다는 행동 인식에 적합하게 최적화된 모델에서 나온 것이어서 행동 탐지에 최적화된 특징을 생성하기 위해 추가되었다.

학습 단계에서, IDN[3]은 최근 현재 시점의 행동 분류를 위한 분류 손실함수, 임베딩 모듈 내에서 행동 탐지를 위한 분류 손실함수와 현재 시점과 서로 다른 행동 클래스인 경우 점점 멀어지게, 같은 행동 클래스인 경우는 가까워지게 만드는 대비 손실함수(Contrastive Loss Function)의 총 3개로 구성

된 통합 손실함수를 이용해 네트워크를 학습시킨다.

4. TFN 모델

IDN[3]과 같은 저자가 제안한 온라인 행동 탐지 모델인 Temporal Filtering Network(TFN)[4]는 입력 비디오 세그먼트 내에서 최신 현재 시점의 행동 클래스를 탐지함에 있어서 세그먼트 내의 각 프레임은 관련성에 따라서 차등을 두어서 모델링에 사용한다는 기본 가정은 동일하다. IDN[3]과의 차이점은 IDN은 IDU라는 온라인 행동 탐지를 위한 RNN 유닛을 새롭게 정의해서 RNN 네트워크를 통해서 현재 행동을 탐지하는데, TFN[4]은 CNN 기반의 필터링 모듈을 통해 비디오 세그먼트 구간 내에서 각 처리 시점과 최신 현재 시점 간의 관계성 스코어 벡터를 예측하고 이를 이용해, 관련된 부분은 강조하고, 관련되지 않은 부분은 억압시키는 기능을 제공한다. 자세히는 시각 특징과 관련성 스코어를 항목별 곱셈(Element-wise Multiplication) 연산을 적용하여 각 시점별 현재 시점과의 관련성 기반으로 가중치를 준 시각 정보 벡터와 원래 시각 정보 벡터를 항목별 덧셈(Element-wise Addition) 연산을 적용하여 획득한 관련성 가중치를 고려한 시각 정보를 만든다. 관련성이 가중치로 적용된 시각 정보를 이용해 현재 행동 클래스를 인식한다(참고 문헌 [4]의 그림 3).

TFN[4]에서는 3가지 서로 다른 필터링 모듈을 제안하는데 그 중에서 인코더-디코더 필터링 모듈이 가장 높은 성능을 보여주었다.

학습 단계에서, TFN[4]은 관련성 스코어 벡터를 학습하기 위해서 예측한 관련성 스코어와 GT 관련성 스코어 간의 L1 거리에 기반한 손실 함수를 사용하고, 분류 모듈에서 행동 분류 성능을 높이기 위해서 분류 손실 함수를 사용한다.

IV. 온라인 행동 탐지 성능

1. 온라인 행동 탐지 데이터셋

TVSeries 데이터셋[5]은 현실 세계에서 흔하게 접하는 30개의 행동 클래스들에 대한 온라인 행동 탐지를 위해 공개된 데이터셋이다. 이 데이터셋은 6개의 인기 있는 TV 시리즈에 대한 27개의 무편집 비디오와 행동 구간에 대한 시간 어노테이션으로 구성되어 있다. 각 비디오는 하나의 에피소드를 포함하는데, 대략 20분에서 40분 길이이다. 27개 비디오는 학습용 13개, 검증용 7개, 테스트용 7개로 나뉜다.

THUMOS-14 데이터셋[6]은 15개의 행동 클래스에 대한 행동 탐지 성능을 평가하기 위해 공개된 데이터셋이다. 3년간 행동 인식과 오프라인 행동 탐지하는 태스크에 대한 THUMOS 챌린지를 위한 데이터셋으로 사용되었고, 그 이후에는 오프라인 방식의 행동 탐지 비디오에서 벤치마크 데이터셋으로 널리 사용되다가, 최근에는 온라인 행동 탐지 비디오 성능 평가에도 사용되고 있다. 무편집 비디오를 검증용으로 1,010개, 테스트용으로 1,574개 제공한다. 행동 구간의 주석을 제공하는 학습용 무편집 비디오를 제공하지 않아서, 대부분 행동 탐지 연구에서 검증용 데이터로 학습하고, 테스트 데이터로 평가하였다.

2. 온라인 행동 탐지 평가 지표

온라인 행동 탐지 방법의 성능은 프레임당(Per-frame) mAP(mean Average Precision)와 mcAP(mean calibrated Average Precision)의 두 가지 평가 지표로 비교된다. 두 지표는 기본적으로 먼저 각 행동 클래스별로 입력 비디오 내의 모든 프레임에 대한 평균 정확도(Average Precision)를 계산한 다음, 모든 클래

스 대해 평균 정확도의 평균을 계산하여 최종 mAP를 계산한다.

가. 프레임당 mAP

각 행동 클래스별 프레임당 평균 정확도는 다음과 같이 계산한다. 각 행동 클래스에 대해서 입력 비디오의 모든 프레임들을 행동 확률의 내림차순으로 정렬시킨다. 정렬된 프레임들의 현재 위치에서의 정확도는 현 위치가 GT상 행동 구간에 포함되지 않은 경우는 0, (포함된 경우에는 현 위치까지의 GT상 행동 구간에 포함된 프레임 수/현위치까지의 총 프레임 수)로 계산한다. 하나의 행동 클래스에 대한 평균 정확도는 이렇게 계산된 각 프레임별 정확도의 합을 GT상 행동 구간에 포함된 프레임들의 수로 나누어 계산한다. 최종적으로는 전체 행동 클래스에 대해 각 행동별 평균 정확도들의 평균으로 mAP를 계산한다.

나. 프레임당 mcAP

기존 프레임당 mcAP 지표는 각 행동 클래스의 정확도가 행동 인스턴스의 길이가 길거나 행동이 빈번하게 출연하여 비디오 내에서 행동 구간에 포함되는 프레임들의 수가 많을수록 상대적으로 성능이 낮아지는 경향을 보이기 때문에 이를 보정하기 위해서 비디오 내에서 행동 구간에 포함되는 프레임 수에 대한 행동 구간에 포함되지 않은 프레임 수의 비율인 w를 이용한 보정 평균 정확도 기반의 mcAP 지표가 제안되었다[5]. 현 시점에서의 보정 정확도는 내림차순으로 정렬된 프레임들을 현재 위치에서의 행동 구간에 포함되는 프레임 수에 대한 행동 구간에 포함되지 않는 프레임 수의 비율인 w를 이용해 수식 (1)과 같이 계산하고, 이를 이용해 보정 평균 정확도 cAP를 수식 (2)와 같이 계산한다. 그리고 전체 행동에 대한 평균으로 mcAP를 구한다.

$$cPrec(i) = \frac{w\,TP(i)}{w\,TP(i) + FP(i)}, \qquad (1)$$

$$cAP_k = \frac{\sum_i cPrec(i)\,1(i)}{N_P}. \qquad (2)$$

3. 온라인 행동 탐지 성능 비교

TVSeries 데이터셋[5]에 대한 온라인 행동 탐지 모델들의 mcAP 지표 기반의 행동 탐지 성능들이 표 1에 기술되어 있다. 사용한 입력은 크게 외양을 반영하는 RGB, 움직임을 반영하는 플로우(Flow), 그리고 이 둘을 다 반영한 Two−Stream, 총 세 가지 종류로 구분된다. 최근 제안된 온라인 행동 탐지 모델들은 기본적으로 TSN[7] 비디오 백본 네트워크를 통해 추출한 Two−Stream 특징을 사용한 성능이 외양이나 움직임 어느 한 특징을 이용한 것보다 큰 차이로 성능 향상을 보였다. 대부분의 온라인 행동 탐지 모델들이 Two−Stream 특징 중에는 ActivityNet−1.3[9]으로 학습된 TSN[7]에서 추출한 특징을 사용하였는데, IDN[3] 모델은 Kinetics 데이

표 1 TVSeries[5]에 대한 행동 탐지 성능 비교

입력	방법		mcAP(%)
RGB	RED [1]		71.2
	2S−FN [8]		72.4
	TRN [2]		75.4
	IDN [3]		76.6
	TFN [4]		79.0
Flow	FV−SVM [5]		74.3
	IDN [3]		80.3
Two−Stream (TSN [11])	RED [1]	ActivityNet [9]	79.2
	TRN [2]		83.7
	IDN [3]		84.7
	TFN [4]		85.0
	IDN [3]	Kinetics [10]	86.1

표 2 THUMOS-14[6]에 대한 행동 탐지성능 비교

입력	방법		mAP(%)
Offline	MultiLSTM [11]		41.3
	CDC [12]		44.4
Online	RED [1]	ActivityNet [9]	45.3
	TRN [2]		47.2
	IDN [3]		50.0
	TFN [4]		55.7
	IDN [3]	Kinetics [10]	60.3

터셋[10]으로 학습된 TSN[7]에서 추출한 특징을 이용해서 지금까지의 SoTA(State-of-the-Arts) 성능을 기록하고 있다.

THUMOS-14 데이터셋[6]에 대한 온라인 행동 탐지 모델들의 mAP 지표 기반의 행동 탐지 성능들이 표 2에 기술되어 있다. THUMOS-14 데이터셋[6]은 오프라인 행동 탐지 모델들의 성능 비교에 사용된 벤치마크 데이터셋이어서 오프라인 행동 탐지 모델들 중에서 프레임당 행동 확률을 제공하는 모델들[11,12]의 mAP들을 성능 비교에 포함하였다. 프레임 단위의 오프라인 행동 탐지 모델들보다 전반적으로 온라인 행동 탐지 모델들의 성능이 높다. 그리고 TVSeries 데이터셋[5]의 성능과 유사한 패턴을 보여준다. 그리고 Kinetics 데이터셋[10]에 대해 학습한 TSN 비디오 백본 네트워크에서 추출한 특징을 이용한 IDN[3]이 THUMOS-14 데이터셋[6]에 대해서 지금까지의 SoTA(State-of-the-Arts) 성능을 기록하고 있다.

V. 결론

CCTV 등의 스트리밍 비디오에서 행동 탐지 결과를 모니터링하기 위해 온라인 행동 탐지 기술에 대한 요구가 증대되고 있다. 그러나 현재까지는 잘 편집된 비디오를 입력으로 하는 행동 인식 기술들과 비디오 전체를 입력으로 해서 행동의 시작과 종료 시점으로 정확히 찾아내어 해당 구간의 행동 클래스를 분류하는 오프라인 방식의 시간적 행동 탐지에 연구가 집중되어 있다. 최근 온라인 행동 탐지에 대해서 관심이 높아지고 있고 최신 연구들이 제안되어, 온라인 행동 탐지 기술들은 부분적으로 관찰된 비디오 세그먼트 입력을 이용해 비디오 세그먼트 내에서 최신 현재 시점의 행동 확률을 제공한다.

온라인 행동 탐지는 오프라인 행동 탐지에 비해 최근 연구되기 시작한 분야로, 2019년까지 비디오 세그먼트 내의 모든 시각 정보를 RNN 유닛을 이용해 시간적으로 축적하여 최신 현재 시점의 행동을 예측하는 수준이었고, 2020년에 들어와서 발전된 기법들이 제안되고 있다. 오프라인 행동 탐지와 유사하게 좀 더 정교한 구간 탐지가 가능한 방법들이 제안될 것으로 예상된다.

용어해설

Ground-Truth(GT) 인공지능 모델을 학습하고 테스트하는 데 사용되는 데이터셋에서 목적하는 태스크에 대한 정답. 이 GT를 이용해서 해당 데이터셋의 학습 데이터를 이용해 모델을 학습시키고, 테스트 데이터를 이용해 학습된 모델의 성능을 평가함

RNN 유닛 RNN을 구성하는 컴포넌트. 시점별 데이터를 입력으로 받아 과거 정보와 현재 정보를 조합하여 현 시점의 상황 정보를 출력하는 역할을 담당함. 대표적인 RNN 유닛으로는 LSTM과 GRU가 있고, 각 태스크별로 다양한 시점별 데이터를 입력으로 사용하여 현재 시점까지의 상황 정보를 축적하는 RNN 유닛들이 제안됨

약어 정리

CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
mAP	mean Average Precision
mcAP	mean calibrated Average Precision
RNN	Recurrent Neural Network

참고문헌

[1] J. Gao, Z. Yang, and R. Nevatia, “Red: Reinforced encoderdecoder networks for action anticipation,” in Proc. Bri. Mach. Vis. Conf. (BMVC), London, UK, Sept. 2017, pp. 92.1-92.11.

[2] M. Xu et al., “Temporal recurrent networks for online action detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, Rep. of Korea, Oct. 2019, pp. 5532-5541.

[3] H. Eun et al., “Learning to discrimiate information for online action detection,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, June 2020, pp. 806-815.

[4] H. Eun et al., “Temporal filtering networks for online action detection,” Pattern Recognit. (PR), vol. 111, Mar. 2021.

[5] R. De Geest et al., “Online action detection,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Oct. 2016, pp. 269-285.

[6] Y.-G. Jiang et al., “Challenge: Action recognition with a large number of classes,” ECCV'14 THUMOS, 2014, http://crcv.ucf.edu/THUMOS14/

[7] L. Wang et al., “Temporal segment networks: Towards good practices for deep action recognition,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, Netherlands, Oct. 2016, pp. 20-36.

[8] R. De Geest and T. Tuytelaars, “Modeling temporal structure with lstm for online action detection,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Lake Tahoe, NV, USA, Mar. 2018, pp. 1549-1557.

[9] F. C. Heilbron et al., “ActivityNet: A large-scale video benchmark for human activity understanding,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, June 2015.

[10] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, July 2017, pp. 4724-4733.

[11] S. Yeung et al., “Every moment counts: Dense detailed labeling of actions in complex videos,” Int. J. Comput. Vis. vol. 126, 2018, pp. 375-389.

[12] Z. Shou et al., “Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, July 2017, pp. 1417-1426.