

# 시간적 행동 탐지 기술 동향

## Trends in Temporal Action Detection in Untrimmed Videos

문진영 (Jinyoung Moon, jymoon@etri.re.kr) 시각지능연구실 책임연구원  
 김형일 (Hyungil Kim, hikim@etri.re.kr) 시각지능연구실 선임연구원  
 박종열 (Jongyoul Park, jongyoul@etri.re.kr) 시각지능연구실 책임연구원/실장

### ABSTRACT

Temporal action detection (TAD) in untrimmed videos is an important but a challenging problem in the field of computer vision and has gathered increasing interest recently. Although most studies on action in videos have addressed action recognition in trimmed videos, TAD methods are required to understand real-world untrimmed videos, including mostly background and some meaningful action instances belonging to multiple action classes. TAD is mainly composed of temporal action localization that generates temporal action proposals, such as single action and action recognition, which classifies action proposals into action classes. However, the task of generating temporal action proposals with accurate temporal boundaries is challenging in TAD. In this paper, we discuss TAD technologies that are considered high performance in terms of representative TAD studies based on deep learning. Further, we investigate evaluation methodologies for TAD, such as benchmark datasets and performance measures, and subsequently compare the performance of the discussed TAD models.

**KEYWORDS** 비디오 행동 탐지, 시간적 행동 탐지, 행동 구간 국지화, 비디오 행동 이해

### 1. 서론

비디오 행동 이해 기술은 그림 1과 같이, 하나의 행동만을 포함하도록 분할 편집된 비디오(Well-trimmed video)에 대해 그 비디오가 표현하는 행동

클래스를 분류하는 행동 인식(Action recognition) 기술과 행동을 포함하지 않는 백그라운드(Background)와 여러 행동 클래스에 속하는 다수의 행동 인스턴스들을 포함하는 일반 무편집 비디오(Untrimmed video)에서 각 행동 인스턴스별로 행동의

\* DOI: <https://doi.org/10.22648/ETRI.2020.J.350303>

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

발생 위치를 추정하고, 행동을 인식하는 행동 탐지(Action detection) 기술로 구분된다.

객체 인식 및 탐지 분야의 발전과 유사하게, 딥러닝 기반 행동 인식 기술들은 전통적 행동 인식 기술의 성능을 큰 차이로 능가하고 있다. 행동 인식 데이터셋인 UCF-101[1]을 기준으로, 최고 87.9%의 정확도를 기록한 수작업 특징(Hand-crafted features)을 이용한 전통적 행동 인식 기술의 성능 [2]을, 2014년에 RGB와 광학 흐름(Optical flow)을 입력으로 사용한 Two-Stream CNN이 근소한 차이인 88.0%로 능가하기 시작했고[3], 2017년에 98.0%의 성능을 보였으며[4], 최근에는 98.4% 이상의 성능을 보여주고 있다[5]. 그러나 대부분의 실세계 비디오들이 행동이 아닌 백그라운드와 여러 행동 클래스의 행동들을 다수 포함하는 무편집 비디오여서, 실세계 비디오에서 행동을 이해하기 위해서는 비디오 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(그림 1). 시공간 행동 탐지 기술의 경우에는 지도 학습(Supervised learning)을 위한 데이터셋

이 행동 인스턴스별로 시간 구간뿐만 아니라 행동 발생 구간에 포함된 각 프레임 내에서 행동이 발생한 공간적 위치를 주석에 포함해야 한다. 이런 이유로 공개된 데이터셋의 용량이 다른 행동 이해를 위한 데이터셋에 비해 상대적으로 작고 공개된 데이터셋들의 수도 적어서 대부분의 행동 탐지 연구들은 주로 현실적인 시간적 행동 탐지 기술에 대해 진행되고 있다. 따라서 본 고에서는 무편집 비디오에서의 시간적 행동 탐지 기술을 다룬다.

본 고에서는 딥러닝 기반의 시간적 행동 탐지 기술의 대표적 연구 사례 분석을 통해 행동 탐지 기술의 발전 흐름을 살펴보고, 시간적 행동 탐지 기술의 성능 평가를 위해 주로 사용된 공개 데이터셋과 최근 연구들의 행동 탐지 성능을 비교한다.

본 고의 II장에서는 행동 탐지 기술을 개괄적으로 설명하고, III장에서는 딥러닝 기반의 시간적 행동 탐지 기술의 중요 연구 사례들을 중심으로 시간적 행동 탐지 기술 동향을 살펴본다. IV장에서는 행동 탐지 네트워크의 학습 및 검증을 위한 데이터셋과 행동 탐지 성능 측정 방법, 그리고 최신 행동 탐지 모델들의 성능을 비교하고, V장에서는 결론을 맺는다.

## II. 행동 탐지 기술 개요

시간적 행동 탐지는 크게 행동이 발생한 시간 구간을 추정하는 시간적 행동 국지화(Temporal action localization) 태스크와 여기에 추정한 시간 구간을 행동 클래스로 분류하는 행동 인식(Action recognition) 태스크로 구성된다.

시간적 행동 탐지는 그림 2와 같이, 행동 탐지를 위한 학습 과정과 학습된 모델을 이용해 행동 탐지 결과를 출력하는 추론 과정으로 나뉜다. 먼저 입력 비디오를 전처리한 후, 비디오 특징을 추출한다.

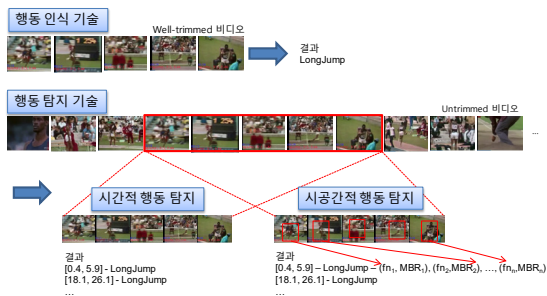


그림 1 행동 이해 기술 개념도

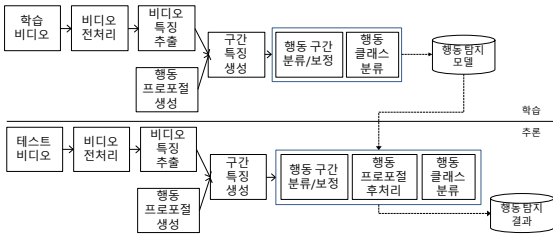


그림 2 일반적인 행동 탐지 파이프라인

그리고 슬라이딩 윈도우 방식으로 비디오 전 구간에 대해서 행동 구간의 후보인 행동 프로포절들을 생성한 다음, 각 행동 프로포절별로 대응하는 행동 구간 특징을 생성한다.

학습 단계에서는 생성한 행동 구간 특징을 이용하여 주어진 행동 프로포절이 행동 구간인지 백그라운드인지를 구분하는 이진 분류 모델 또는 행동 프로포절에 대한 신뢰도를 스코어링하는 회귀 모델을 학습한다. 행동 프로포절의 신뢰도는 다양하게 정의할 수 있는데, 기본적으로 Ground-Truth(GT) 행동 구간과 행동 프로포절 간의 temporal Intersection over Union(tIoU) 값에 기반한다. 추가적으로, 보다 정교한 행동 구간을 획득하기 위해 행동 프로포절의 바운더리(Boundary)를 보정하는 오프셋을 예측하는 회귀 모델을 학습하기도 한다.

여기까지가 행동 국지화 기술이고, 행동 탐지 기술은 탐지된 행동 구간에 대해 행동 클래스를 인식하는 분류 모델을 학습한다.

행동 프로포절의 분류 및 보정과 행동 클래스 분류 태스크는 하나의 동일한 네트워크 또는 서로 다른 복수 개의 네트워크들로 구현될 수 있다. 이 네트워크 학습을 위해서 각 태스크별 손실 함수들을 개별적으로 정의하고, 각각 최적화에 이용하거나 개별 손실 함수들을 결합한 전체 손실 함수를 정의하여 멀티태스킹 최적화에 이용한다.

추론 단계에서는 다수의 행동 프로포절들을 생성하고, 학습된 모델을 이용하여 입력 행동 프로포절이 행동인지 또는 백그라운드인지를 판단하고, 행동인 경우 추론된 행동 프로포절의 보정 오프셋을 이용하여 행동 구간의 시작과 종료 시점을 보정하여 행동 바운더리를 정교화한 뒤 NMS(Non-Maximum Suppression)로 행동 구간을 후처리한다. 시간적 NMS 알고리즘은 임계치를 초과하여 시간적으로 중첩된 여러 행동 프로포절들 중에서 신뢰도가 높은 하나만 남겨서 중첩된 행동 프로포절들을 제거한다. 따라서 탐지된 행동 프로포절들 수에 대한 제약이 있을 때, NMS를 거치면 높은 신뢰도의 중첩되지 않은 행동 프로포절들이 탐지 결과에 포함되어, 재현율이 높은 탐지 결과를 반환할 수 있다. 마지막으로 행동 클래스 분류 모델을 이용하여 후처리를 완료한 행동 프로포절에 대해 행동 인식을 수행한다.

### III. 딥러닝 기반 행동 탐지 기술동향

본 절에서는 중요 딥러닝 기반의 시간적 행동 탐지 및 국지화 기술 사례들을 분석하고, 행동 탐지 기술의 전체적 발전 흐름을 살펴본다.

#### 1. 초기 행동 탐지 모델

초기 딥러닝 기반 행동 탐지 모델들[6-8]은 시간적 행동 탐지 챌린지인 THUMOS[9]를 위해 제안되었는데, 수작업 특징(Hand-crafted features) 또는 부분적으로 CNN으로 추출한 특징을 이용하고, 분류기로는 SVM을 사용하였다. 최초의 딥러닝 기반의 행동 인식 모델인 Two-Stream CNN[4]이 발표되기 이전에 제안된 모델들이어서, 단순 특징 추출을 위해 딥 뉴럴 네트워크를 이용하여 완전한 딥

러닝 기반 행동 탐지 모델이라고 보기는 어렵다.

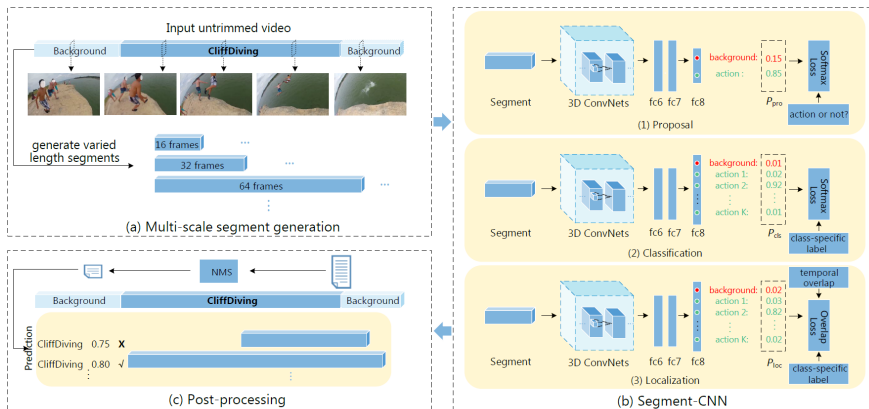
뒤이어 RNN 기반의 행동 탐지 모델들[10–12]이 제안되었는데, 이들은 시계열 데이터인 비디오를 입력으로 행동의 시간적 모델링에 적합한 LSTM 등의 RNN을 이용해 각 프레임 또는 프레임 시퀀스별로 행동 클래스별 확률을 예측하고, 같은 행동 클래스에 속하는 연이은 또는 근접 이웃 프레임들을 휴리스틱 방법으로 병합하여 행동 구간을 획득하였다. 그러나 추정된 행동 클래스별 확률은 비슷한 행동들 간에 혼동될 수 있는데, 이로 인해 하나의 행동 구간 내에서 서로 다른 클래스로 잘못 분류되면 다른 구간으로 인지되어 실제보다 과도하게 끊어진 행동 구간이 생성될 수 있다. 그리고 행동 구간의 신뢰도는 행동 클래스별 확률을 전체 구간에서 평균값으로 산정하여 정확도가 낮았다.

## 2. 2단계 행동 탐지 모델

프로포절 생성과 분류의 2단계로 구성된 객체 탐지 접근 방법들[13–15]의 큰 성공에 영향을 받아, 행동 탐지 모델들도 행동 프로포절 생성과 분류를 구분하거나 행동 프로포절 생성에 집중한 행

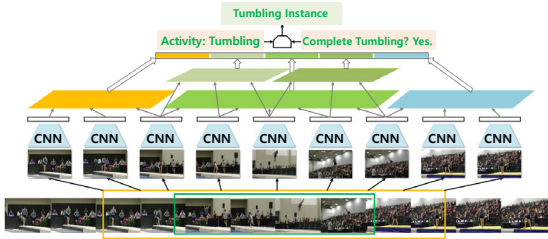
동 탐지 모델들이 제안되기 시작했다[16–23]. 2단계 행동 탐지 모델들과 초기 행동 탐지 모델들과의 가장 큰 차이는 프레임이 아닌 행동 프로포절 자체가 행동 구간인지 백그라운드인지를 판단하는 것이다.

최초의 2단계 시간적 행동 탐지 모델인 S-CNN[16]은 그림 3과 같이 프로포절, 분류, 국지화 네트워크를 포함한 총 3개의 C3D[17]를 베이스로 하는 3D CNN으로 구성된다. 먼저 다중 스케일의 슬라이딩 윈도우 방식으로 시작과 종료 시점이 확정되는 가변 길이의 행동 세그먼트들을 생성한다. 그리고 프로포절 네트워크에서 입력 행동 세그먼트가 행동 구간인지 백그라운드인지 이진 분류를 수행한다. 분류 네트워크는 행동 세그먼트의 행동 클래스 확률을 예측하는 모델로 학습 단계에서 학습된 모델은 국지화 네트워크의 초기화에 이용하고, 추론 단계에서는 사용되지 않는다. 국지화 네트워크는 클래스 분류와 중첩을 모두 고려한 손실 함수를 이용하여, GT와 많이 중첩된 행동 세그먼트의 행동 클래스 확률을 높이도록 중첩 정도를 고려한 행동 클래스 확률을 출력한다. 그리고 후처리로 NMS를 통해 중복된 세그먼트들을 제거한다.



출처 Reprinted with Author's Permission from <https://arxiv.org/abs/1601.02129>

그림 3 S-CNN 구성도



출처 Reprinted with Author's Permission from <https://arxiv.org/abs/1704.06228>

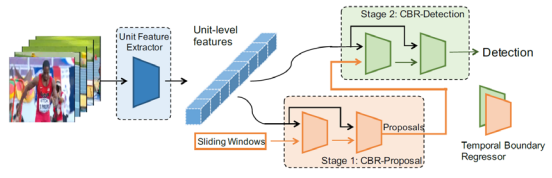
그림 4 SSN 개념도

행동 구간 이외에 행동 구간의 시작 전과 종료 후를 행동 컨텍스트로 고려한 행동 탐지 모델인 SSN[18]은 객체 인식의 공간적 피라미드 풀링[19]에 영향을 받아, 그림 4와 같이 행동 구간을 행동의 시작 전, 진행 중, 종료 후 3단계로 나누고, 진행 단계는 좀 더 세분화하고 구조적으로 피라미드로 쌓아올리는 구조화된 시간적 피라미드 풀링 기법을 제안하였다.

SSN[18]은 이 풀링 기법을 이용하여 해당 프로포절의 행동을 인식하기 위해서 진행 구간에 대해서만 입력으로 사용하는 1개의 행동 분류기와 전 구간을 고려한 각 행동 클래스별 K개의 완료/미완료 분류기로 구성되었다. 그리고 딥러닝 기반 행동 탐지 모델 최초로 행동 구간의 바운더리를 교정하는 보정 오프셋을 예측하는 회귀 모델을 포함하였다.

또, SSN[18]은 객체 탐지 방법들[13-15]에서 사용된 objectness와 유사한 개념의 actionness라는 스코어를 이용하여 임계치를 넘는 actionness를 가지는 행동 구간에서 연속된 또는 근접한 프레임 시퀀스를 그룹핑하여 행동 구간을 생성하는 TAG (Temporal Actionness Grouping) 방법을 제안하였다. TAG는 추후 스코어 기반으로 행동 구간을 생성할 때 다른 연구들에서 많이 사용되었다.

CBR[20]은 슬라이딩 윈도우 방식으로 프로포절



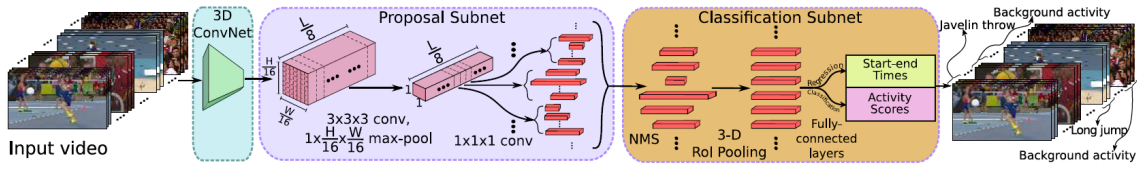
출처 Reprinted with Author's Permission from <https://arxiv.org/abs/1705.01180>

그림 5 CBR 구성도

생성 시 행동 구간의 시작과 종료 시점이 부정확한 문제를 해결하고자, 보정 오프셋을 예측하는 모듈을 여러 번 반복하는 캐스케이드 방식을 제안하였다. CBR[20]은 그림 5와 같이, CBR-Proposal 모듈에서 획득한 보정 오프셋 예측치를 이용하여 바운더리를 수정한 프로포절을 CBR-Detection 모듈에서 행동 인식에 사용한다. CBR[20]의 저자들은 바운더리 부정확성 문제를 해결하는 행동 국지화 모델인 TURN[21]도 제안하였는데, 참고문헌 [20,21]에서 특정 수의 프레임 시퀀스인 유닛 단위로 비디오 특징을 추출하고, 프로포절 구간과 정해진 유닛 수의 컨텍스트 구간 각각의 평균값 풀링 결과를 이어 붙여 프로포절 특징으로 이용하였다.

다중 스케일의 슬라이딩 윈도우 기반의 프로포절 생성 방식이 윈도우 스케일별로 입력 영상을 여러 번 처리해야 하는 단점이 있어서, 이를 보완하기 위해 다중 스케일 앵커 기반의 단일 패스 프로포절 생성 기법을 사용하는 행동 탐지 및 국지화 모델들이 제안되었다[22,23].

R-C3D[22]은 그림 6과 같이, RGB 프레임들을 입력으로 C3D[17]를 베이스 모듈로 기본 특징을 만들어서 각 위치별 다중 스케일의 앵커에 대응하는 프로포절 특징을 만들고 프로포절 서브 네트워크에서 백그라운드인지 행동인지 분류한다. 그 다음 분류 서브 네트워크에서 3D RoI 풀링 방법으로 서로 다른 크기의 프로포절에 대해서 동일한 프로



출처 Reprinted with Author's Permission from <https://arxiv.org/abs/1703.07814>

그림 6 R-C3D 구성도

포절 특징을 만든 뒤, 행동 구간의 시작과 종료 시점을 보정하는 오프셋을 예측하고 행동 클래스를 분류한다.

TCN[23]은 SSN[18]과 유사하게, 행동 시작 전과 종료 후 구간을 행동 컨텍스트라고 명명하고, 행동 구간만 입력으로 이용해 행동 클래스를 분류하고 행동 진행 구간에 행동 컨텍스트를 포함해서 행동인지 백그라운드인지를 구분하는 행동 탐지 네트워크를 제안했다.

2단계 행동 탐지 모델들이 행동 프로포절이 행동인지 백그라운드인지를 판단하고, 행동 진행 구간뿐만 아니라 행동 시작 전과 종료 후의 컨텍스트를 고려하며, 비디오를 프레임의 시퀀스인 유닛 단위로 나누어 비디오 특징을 뽑아 사용하는 방식들은 다음 연구들에 많은 영향을 미쳤다.

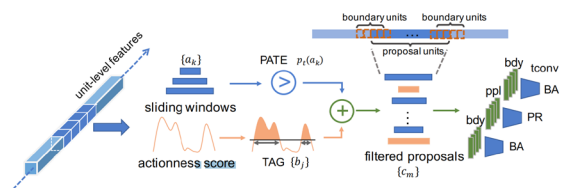
### 3. 스코어 기반 프로포절 생성 모델

슬라이딩 윈도우 방식으로 생성한 행동 프로포절들은 전 구간에서 고르게 생성되므로 전역적 측면에서 미탐지되는 행동의 수는 줄여 주지만, actionness 스코어를 그룹핑해서 생성한 행동 프로포절들에 비해 시작 및 종료 시점의 정확도는 떨어진다. 그러나 스코어 기반 프로포절들은 스코어의 정확도가 높지 않으면 미탐지되는 행동의 수가 많아진다. 이 문제를 해결하기 위해 스코어 기반으로 프로포절들을 생성하는 모델들이 제안

되었다[24–29].

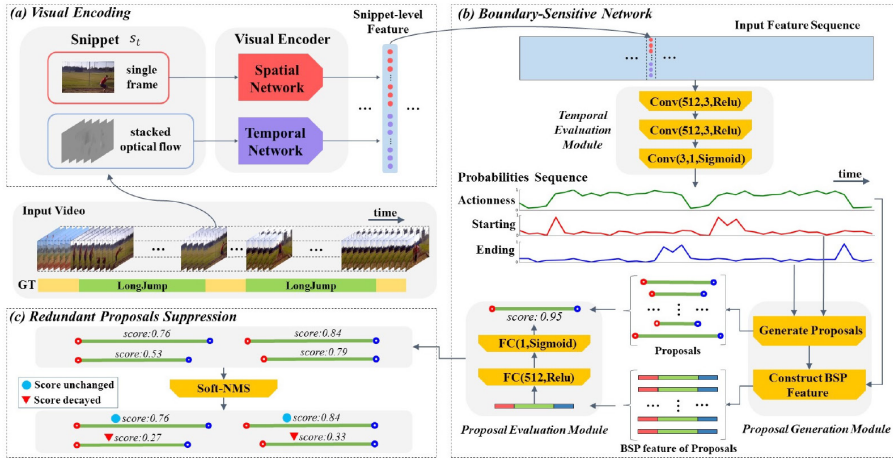
CTAP[24]에서는 각 방식의 단점을 상호보완하기 위해 이 둘을 결합시킨 하이브리드 방식을 제안하였다. 그림 7과 같이, 슬라이딩 윈도우 기반 프로포절들과 actionness 스코어 기반 프로포절들을 함께 사용하되, 정확도가 떨어지는 슬라이딩 윈도우 방식의 프로포절들을 필터링하여 제거시킨다. 이를 위해 슬라이딩 윈도우 기반 프로포절의 actionness 스코어가 얼마나 믿을만한지 예측하는 Proposal-level Actionness Trustworthiness Estimator(PATE)를 학습시킨다.

추론 단계에서는 슬라이딩 윈도우 방식으로 생성된 프로포절들 중에서 특정 임계치보다 작은, 즉 actionness를 믿을 수 없는 프로포절들을 제거하는 상보적 필터링(Complementary filtering)을 수행한다. 그리고 이렇게 생성된 프로포절들에 대해 프로포절 유닛과 행동의 시작과 종료 구간을 포함하는 바운더리 유닛 각각을 시간적 컨볼루션 레이어를 이용해 행동을 시간적으로 모델링한 뒤, tIoU에 기



출처 Reprinted with Author's Permission from <https://arxiv.org/abs/1807.04821>

그림 7 CTAP 개념도



출처 Reprinted with Author's Permission from <https://arxiv.org/abs/1806.02964>

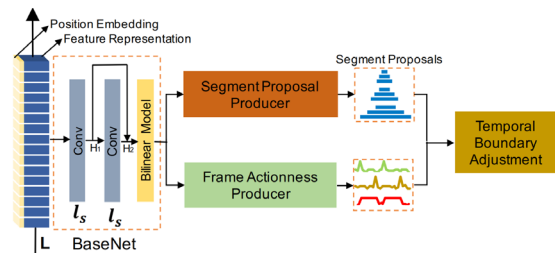
그림 8 BSN 구성도

반한 프로포절 랭킹 스코어와 시작과 종료 시점의 보정 오프셋을 예측한다. 이 예측한 보정 오프셋을 반영해 프로포절을 보정한 후 NMS로 중첩된 프로포절들을 제거한다. 그리고, 랭킹 스코어를 기반으로 스코어가 높은 N개의 보정된 행동 프로포절들을 행동 국지화 결과로 반환한다.

BSN[25]은 처음으로 슬라이딩 윈도우 방식에서 탈피하여, 완전한 스코어 기반 방식으로 높은 성능을 보여준 행동 국지화 모델이다. 슬라이딩 윈도우 방식이 비디오 전 구간을 빠짐없이 커버할 수 있지만, 행동에 대한 시각 정보를 활용하지 않아서 비효율적인데, BSN[25]은 이를 극복하기 위해서 그림 8과 같이 actionness뿐만 아니라 시작과 종료를 포함해 총 3종의 행동 스코어 예측 결과를 이용해서 행동 관련 시각 정보를 기반으로 행동 프로포절들을 생성한다. 특정 임계치를 넘는 높은 값들이나 임계치를 넘지 않는 피크의 시작과 종료 시점들을 조합하여 초기 프로포절들을 생성한다. 그리고 이 초기 프로포절의 진행 구간과 바운더리 구간 각각의 actionness 시퀀스를 선형 보간법을 통해서 16개와 8개로 길이를 맞춘 뒤에 이어 붙여서 바운더

리에 민감한 프로포절 특징을 생성한다. 이 프로포절 특징을 이용해서 입력 프로포절의 스코어와 바운더리 보정 오프셋을 예측하는 프로포절 평가 모듈을 이용해 개선된 프로포절과 프로포절 스코어를 반환한다.

MGG[26]는 기본적으로 앵커 방식으로 프로포절들을 생성하고, 이 프로포절들의 바운더리를 정교하게 보정을 위해 actionness, 시작, 종료의 총 3종 행동 스코어를 활용한다. MGG[26]는 그림 9와 같이 비디오 시퀀스에 대해 일반적인 비디오 특징과 위치 정보를 결합시킨 프로포절 특징으



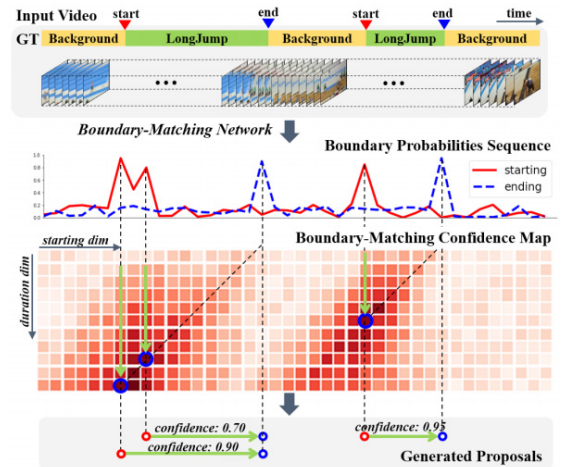
출처 Reprinted with Author's Permission from <https://arxiv.org/abs/1811.11524>, CC BY-NC-ND

그림 9 MGG 구성도

로 사용할 특징을 생성하는 BaseNet, 앵커 기반의 행동 프로포절들을 생성하는 Segment Proposal Producer, 스코어 기반 행동 프로포절들을 생성하는 Frame Actionness Producer, 그리고 행동 프로포절의 바운더리를 정교화시키는 Temporal Boundary Adjustment의 4개 모듈로 구성된다. Segment Proposal Producer 모듈에서는 앵커 방식으로 생성된 프로포절들이 행동을 포함하는지 아닌지를 이진 분류하고, 바운더리 보정 오프셋을 예측하도록 학습시켜 세그먼트 프로포절들을 생성한다. Frame Actionness Producer 모듈에서는 시작, 중간, 종료의 행동 스코어 3종을 예측한다. Temporal Boundary Adjustment 모듈에서는 먼저 NMS로 중첩된 프로포절들을 제거한 뒤, 2단계에 걸쳐 행동 프로포절의 바운더리를 보정한다. 1단계에서 세그먼트 기반으로 탐지된 행동 프로포절들에 대해서, 전체 행동 길이에 비례하게 시작 시점을 중심으로 하는 시작 영역과 종료 시점을 중심으로 하는 종료 영역을 만들고, 영역 내에서 시작 및 종료 스코어의 최대값이 특정 임계치를 넘으면, 가장 높은 지점으로 시작 및 종료 시점을 변경한다. 2단계에서 세그먼트 프로포절들이 actionness 기반의 프로포절들과 임계치 이상이 중첩되면, 기존 세그먼트 프로포절들을 더 정확한 시작 및 종료 시점으로 보정하기 위하여 actionness 기반 프로포절들로 교체한다.

기존 스코어 기반 프로포절 생성 방법들이 1차원의 스코어 시퀀스를 예측[24–26]한 것과 대조적으로, 행동 스코어의 정확도를 높이기 위해 조밀하게 분포된 행동 프로포절들을 나타내는 2차원 맵으로 예측한 방법들이 제안되었다[27–29].

BMN[27]은 BSN[25]의 후속 연구로, 행동 프로포절 생성은 BSN[25]과 동일하게 1차원 시퀀스로 예측한 시작 및 종료 시퀀스를 이용하고, 그림 10



출처 Reprinted with Author's Permission from <https://arxiv.org/abs/1907.09702>

그림 10 BMN 개념도

과 같이 생성된 행동 프로포절의 스코어를 2차원의 신뢰도 맵으로 예측한다. 이 맵에서 x축은 행동 시작 위치를 y축은 행동의 길이를 나타내어, 각 셀이 하나의 고유한 프로포절을 나타내어 조밀하게 분포된 프로포절에 대해서 학습시켜 정확도 높은 스코어를 획득할 수 있다.

SRG[28]는 actionness와 유사하지만, 레퍼런스 유닛이 포함한 행동 인스턴스와 관련이 있는지를 나타내는 relatedness라는 새로운 행동 스코어를 제안하고, relatedness와 행동의 종료와 시작을 포함하는 총 3종의 행동 스코어 모두를 2차원의 맵으로 예측하였다. 정해진 사이즈의 윈도우를 유닛 1개 단위로 밀어서 전 구간을 표현한 2차원 맵을 통해 조밀하게 분포한 구간에 대한 스코어를 학습시켜 정확한 스코어를 획득하고자 했다. 예측한 스코어 3종을 이용해 다음의 2가지 방법을 통해 초기 행동 구간을 생성한다. 특정 임계치를 넘는 relatedness를 그룹핑해서 첫 번째 타입의 초기 행동 구간을 생성한다. 그리고 최고점 또는 피크인 시작과 종료 시점을 이용해 해당 구간의 값만 통과시키는 이진 웨이



트 시퀀스를 만들어서 relatedness 시퀀스에 가중치 합을 한 뒤에 특정 임계치를 넘는 relatedness를 그룹핑해서 두 번째 타입의 초기 행동 구간을 생성한다. 생성한 모든 초기 행동 구간들은 행동 구간의 신뢰도와 보정 오프셋을 예측하는 행동 구간 평가 모듈을 거쳐 보정된 행동 프로포절들을 결과물로 반환한다.

SRG[28]은 relatedness라고 해서 행동 인스턴스별 관련성을 나타내는 스코어를 사용하므로, 동일한 복수의 행동들이 반복되어도 행동별로 정확하게 프로포절 바운더리를 구분하는 능력을 보여주나, 2차원 맵상의 하나의 셀이 고유한 행동 구간을 나타내지 않기 때문에 행동 구간이 중첩되는 경우에는 제대로 행동 구간 표현이 불가하다.

DBG[28]는 행동의 시작, 종료, 완료의 총 3종 스코어를 조밀하게 분포한 프로포절을 나타내는 2차원 맵 형태로 예측하여 프로포절을 생성한다. x축은 시작 위치를 y축은 종료 위치를 나타내어, 2차원 맵에서 하나의 셀은 하나의 고유한 행동 프로포절을 나타낸다. 행동 프로포절의 바운더리에 특징을 생성하기 위해서 Dual Stream Base 모듈에서 RGB 스트림, 광학 플로우 스트림 2개를 병합한 3개의 스트림을 평균하여 행동 스코어 특징을, 그리고 두 스트림을 요소별 합으로 병합한 듀얼 스트림 특징을 생성한다. 두 종류의 기본 특징으로 프로포절 특징을 생성하고, 이를 이용해 Action-aware Completeness Regression 모듈에서 2차원 맵 형태의 행동 완료 스코어를, Temporal Boundary Classification 모듈에서 시작과 종료 스코어를 예측하는 모델을 제안하였다.

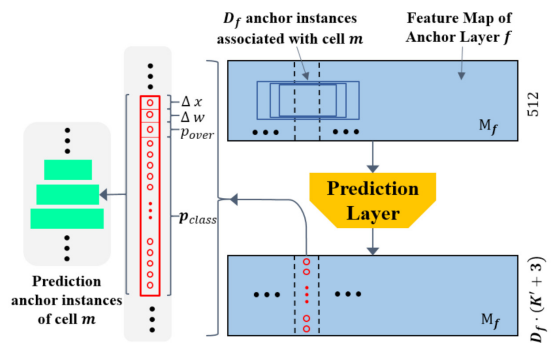
1차원 스코어 시퀀스를 이용해 행동 구간을 생성하는 것은 스코어별 임계치를 지정하거나, 휴리스틱 방법으로 스코어 시퀀스를 이용해 프로포절 생성할 때 최적화의 문제가 있다. 이에 반해 2차

원 스코어 맵 기반으로 행동 프로포절들을 생성하는 행동 국지화 방법들은 맵을 구성하는 각 셀들이 고유한 행동 구간을 나타내는 경우, 프로포절 신뢰도 맵에서 높은 스코어 상위 K개를 뽑아서 행동 프로포절 탐지 결과로 반환하면 되기 때문에, 직관적인 방법으로 최적화된 프로포절 생성이 가능하다.

### 4. Single-Shot 행동 탐지 모델

SSD[30]와 같은 single-shot 객체 탐지 방법들의 영향을 받아, 프로포절 생성 단계를 제거하고 단일 네트워크 안에서 행동 구간과 신뢰도를 예측하는 single-shot 행동 탐지 방법들도 제안되었다 [31,32].

SSAD[31]은 프레임 시퀀스인 유닛별로 RGB, 광학 흐름, C3D[17] 기반의 행동 분류기를 통해 획득한 행동 클래스 확률 3개를 이어 붙여 유닛별 액션 스코어 특징을 생성한 뒤, 수용 영역(Receptive field)이 서로 다른 앵커 레이어별로 각 위치에서 특징 맵을 만든다. 그리고 그림 11과 같이 앵커 레이어별 특징 맵을 입력으로 예측 레이어에서는 각 위치에서 생성된 다중 스케일의 모



출처 Reprinted with Permission from <https://arxiv.org/abs/1710.06236>

그림 11 SSAD Prediction Layer 동작 개념도

든 앵커들에 대해서 보정 오프셋, 프로포절 신뢰도, 행동 클래스별 확률을 모두 포함한 예측 결과를 반환한다.

S3D[32]는 앵커 방식으로 생성한 전체 252개의 행동 프로포절들에 대해서 single-shot으로 행동을 탐지한다. 입력 비디오는 C3D[17]를 기반으로 하는 전 구간에 대한 기본 특징을 생성하고, 5개의 컨볼루션 레이어로 구성된 보조적 시간적 특징 레이어를 통해 다른 사이즈의 수용 영역을 가지는 프로포절 특징 맵을 추출하며, 각 특징 맵별로 생성된 프로포절 특징을 이용하여 다중 앵커에 대해 single-shot 행동 탐지를 수행한다.

사용할 프로포절의 수를 조절할 수 있는 스코어 기반 행동 모델과 비교해서 single-shot 행동 탐지 모델들은 전 구간에 고르게 분포시킨 정해진 수의 프로포절들에 대해서만 행동 탐지가 가능하다. 그래서 일반적으로 스코어 기반 행동 탐지에 비해서 탐지 성능이 낮고, 속도 측면에서 우수성이 검증되지 않아서 single-shot 행동 탐지 모델들은 객체 탐지 분야와는 달리, 행동 탐지 분야에서는 활발한 연구가 이루어지지 않았다.

## 5. GCN 기반 행동 탐지 모델

노드와 노드들 간의 관계를 나타내는 에지로 구성된 그래프에 대한 연산을 처리하는 Graph Convolutional Network(GCN)[33]를 이용한 스켈레톤 기반의 행동 인식[34]과 비디오 분류 연구[35]들의 영향을 받아, GCN을 행동 탐지에 도입하여 행동 구간을 정교하게 하는 모델들이 제안되었다[35,36].

P-GCN[36]은 중첩된 프로포절들과 어느 범위 내에 존재하는 프로포절들은 관련성이 있다고 가정하고, 프로포절은 노드로, 중첩된 프로포절은

상황정보 관계를 나타내는 에지로, 중첩되지 않았지만 어느 범위 내에 근접하여 위치한 프로포절들은 주변 관계를 나타내는 에지로 그래프를 정의하였다. 프로포절 구간 내의 정보만 이용한 프로포절 특징을 이용해 첫 번째 GCN에서 행동 카테고리를 예측하고, 컨텍스트를 포함해 확장된 프로포절 특징을 이용하는 두 번째 GCN에서 보정 오프셋을 예측한다.

그러나 GCN에서 프로포절 수가 증가하면 연산 비용이 증가하므로, P-GCN[36]에서는 효과적인 학습을 위해 가중치 업데이트 시에 모든 프로포절이 아닌 샘플링된 인접 프로포절들로부터만 정보를 통합한다. 이런 방법들을 쓴다고 해도 GCN에서 처리할 수 있는 노드 수의 한계가 있어 P-GCN[36]에서는 BSN[25]에서 생성한 프로포절들을 사용한다.

따라서 GCN 기반 행동 탐지 및 국지화 모델들은 초기 프로포절 생성부터 시작하는 독자적인 알고리즘으로는 힘들고, 단순히 보정 오프셋을 예측하는 것에 비해 한층 수준 높은 보정 모듈이 될 것으로 예상된다.

## IV. 행동 탐지 성능 비교

본 절에서는 행동 탐지 및 국지화 기술의 성능을 비교하기 위해서 사용되고 있는 행동 탐지 데이터셋들에 대해 살펴본 뒤, 행동 탐지 및 국지화 기술을 질적 수준을 평가하는 데 사용되는 평가 방법을 알아보고, 각 행동 탐지 데이터셋별로 앞에서 살펴본 행동 탐지 및 국지화 모델별 성능을 비교한다.

### 1. 행동 탐지 데이터셋

시간적 행동 탐지 기술은 행동 탐지 챌린지[9,38]

에서 제공하는 데이터셋과 챌린지에서 제공한 성능평가 방법을 그대로 도입하여, 비교적 벤치마킹 데이터셋과 성능평가 방법이 일찍 정리된 분야이다. 2017년 이후로는 대부분의 메이저 학회 및 저널에 공개되는 모델들은 공통적으로 THUMOS-14[9]와 ActivityNet-1.3[39] 데이터셋에 대해 성능 평가를 수행하였다.

15개 행동 클래스를 다루는 THUMOS-14[9]는 행동 탐지 관련해서는 무편집 비디오를 검증용으로 1,010개, 테스트용으로 1,574개 제공하는데 행동 구간의 주석을 제공하는 학습용 무편집 비디오를 제공하지 않아서 대부분 연구에서 검증용 데이터로 학습하고, 테스트 데이터로 평가하였다.

ActivityNet-1.3[39]은 최초의 대용량 행동 탐지 데이터셋으로, 200 행동 클래스에 대해 총 10K 학습 비디오와 5K 검증 비디오, 5K 테스트 비디오를 제공한다. 테스트 데이터에 대한 정답을 공개하지 않고, 챌린지 평가 서버를 통해서만 성능을 측정할 수 있어서 대부분 검증 비디오로 성능을 평가한다.

THUMOS-14[9]가 ActivityNet-1.3[39]에 비해 데이터셋에 포함된 비디오 수는 적지만, THUMOS-14[9]는 한 비디오당 평균 15개 이상의 행동 인스턴스들을 포함하고, 비디오 내의 71%를 백그라운드가 차지한다. 그에 비해 ActivityNet-1.3[39]은 비디오당 평균 1.65개의 행동을 포함하고, 대부분의 행동이 중간 지점을 중심으로 비디오의 70% 정도 차지한다. 그래서 ActivityNet-1.3[39]이 행동 탐지에 사용하기에는 한 비디오에 포함된 행동 수가 적다는 의견이 있다. THUMOS-14[9]는 데이터셋 크기가 작기는 하지만, 행동 탐지 태스크를 수행하기에 도전적인 데이터셋이어서, 현재 두 데이터셋 모두 동등한 비중으로 행동 탐지 성능 비교에 사용되고 있다.

## 2. 행동 탐지 평가 방법

행동 국지화 기술의 성능 평가를 위해서 특정 구간의 tIoU 값에 대해 주어진 평균 프로포절 수만큼의 프로포절들을 이용하여 측정한 average recall 값인 average recall@average number(AR@AN)을 주로 사용한다. 사용한 프로포절 수가 많을수록 재현율(recall)이 올라가기 때문에 사용한 비디오당 평균 프로포절 수인 AN에 따라 값을 비교한다. tIoU는 탐지한 행동 구간이 GT와 매칭하는지 안 하는지를 평가하는 기준이 되는데, 평균을 내는 tIoU 구간은 최소 0.5 이상의 tIoU에 대해서 매칭한다고 인정하지만 챌린지별로 근소한 차이가 나는 구간을 지정하고 있다. 그리고 또 다른 평가 기준으로 AR과 AN 사이의 the area under the curve(AUC)를 이용한다.

행동 탐지 기술의 성능 평가를 위해서 THUMOS-14[9]에 대해서는 각 tIoU별로 행동 클래스별 AP를 평균하여 계산한 mean Average Precision(mAP@tIoU)을 주로 사용하고, ActivityNet-1.3[39]에 대해서는 AR@AN=100과 AUC를 주로 이용한다.

행동 탐지 성능의 평가를 위해서 행동 국지화 모델들은 AR@AN 또는 S-CNN의 분류 네트워크[16]나 UntrimmedNet[40] 등의 외부 공개된 행동 인식기를 붙여서 측정한 mAP@tIoU로, 행동 탐지 모델들은 자체 행동 클래스 분류기에서 획득한 mAP@tIoU로 정량적 성능을 제시한다.

## 3. 행동 탐지 성능 비교

THUMOS-14[9] 데이터셋에 대한 행동 프로포절 생성과 관련된 행동 국지화 성능은 표 1과 같이 AN별 AR 비교한다.

표 1 THUMOS-14[9] 국지화 성능(AR@AN)

모델	@50	@100	@200	@500
SCNN prop[16]	17.22	26.17	37.01	51.57
TURN[21]	21.86	31.89	43.02	57.63
CTAP[24]	32.49	42.61	51.97	-
BSN[25]	37.46	46.06	53.21	61.35
MGG[26]	39.93	47.75	54.65	61.36
BMN[27]	39.36	47.72	54.84	62.19
SRG[28]	42.19	49.72	56.71	63.78
DBG[29]	40.89	49.24	55.76	62.21

그리고, ActivityNet-1.3[39] 데이터셋에 대한 국지화 성능은 표 2와 같이, AN=100일 때의 AR과 AUC로 비교한다.

THUMOS-14[9]가 ActivityNet-1.3[39]보다 전반적으로 성능이 낮지만, 새로운 기술이 제안됨에 따라 큰 폭으로 성능이 향상되고 있고, ActivityNet-1.3[39]은 성능이 높은 수준에서 시작했지만 미미한 성능 향상을 보여주며 발전하고 있다. 이는 비디오당 들어 있는 행동의 수와 데이터셋의 크기 차이 등의 데이터셋의 특성 차이로 발생하는 것으로 보고 있다. 그리고 행동 탐지 모델들 중에서 최근 제안된 2차원 맵을 사용하는 스코어 기반의 모델들이 전반적으로 높은 국지화 성능을 보여준다.

그리고 행동 탐지 성능은 표 3과 4처럼, 데이터

표 2 ActivityNet-1.3[39] 국지화 성능(AR@AN)

모델	AR@AN=100	AUC
CTAP[24]	73.17	65.72
BSN[25]	74.16	66.17
MGG[26]	74.56	66.54
BMN[27]	75.01	67.10
SRG[28]	74.65	66.06
DBG[29]	76.65	68.28

표 3 THUMOS-14[9] 행동 탐지 성능(mAP@tIoU)

모델	@0.3	@0.4	@0.5
S-CNN[16]	36.3	28.7	19.0
SSN[18]	51.9	41.0	29.8
TCN[23]	-	33.3	25.6
R-C3D[22]	44.8	35.6	28.9
TURN[21]	44.1	34.9	25.6
CBR[20]	50.1	41.3	31.0
SSAD[31]	43.0	35.0	24.6
S3D[32]	47.9	41.2	32.6
BSN[25]	53.5	45.0	36.9
MGG[26]	53.9	46.8	37.4
SRG[28]	54.5	46.9	28.4
BMN[27]	56.0	47.4	38.8
DBG[29]	57.8	49.4	39.8
P-GCN[36]	63.6	57.8	49.1

셋별로 정해진 tIoU별 mAP로 비교한다.

외부 행동 인식기를 연동한 스코어 기반의 행동 국지화 모델의 성능이 대체로 행동 탐지 모델의 성능보다 더 나은 성능을 기록하고 있다. 이는 행동 구간 국지화와 인식을 모두 포함하는 종단 간 최적화된 행동 탐지 모델을 고수하기보다는, 행동 국지화 모델을 통해 정교한 바운더리를 가지는 행동 프로포절들을 잘 생성하는 것이 성능을 좌우하는 핵심 요소이고, 바운더리가 정확한 행동 프로포절은 외부 공개된 행동 인식기를 사용해도 좋은 인식 성

표 4 ActivityNet-1.3[39] 행동 탐지 성능(mAP@tIoU)

모델	@0.5	@0.75	@0.95	평균
S-CNN[16]	45.30	26.00	0.20	23.80
R-C3D[22]	26.80	-	-	-
TCN[23]	36.44	21.15	3.90	-
SSN[18]	39.12	23.48	5.49	23.98
BSN[25]	46.45	29.96	8.02	30.03
SRG[28]	46.53	29.98	4.83	29.72
BMN[27]	50.07	34.78	8.29	33.85
P-GCN[36]	48.26	33.16	3.27	31.11

능을 얻을 수 있음을 보여준다.

## V. 결론

본 고에서는 실세계 비디오를 이해하는 데 필수적인 시간적 행동 탐지 기술에 대해 최근 연구된 딥러닝 기반 행동 탐지 모델을 중심으로 기술 동향을 살펴보았다.

시간적 행동 탐지 기술은 THUMOS[9]와 ActivityNet[38]를 포함하는 행동 탐지 챌린지를 통해 시작되어, 행동 탐지 및 국지화 기술은 초기부터 문제 정의가 잘 정리되고, 행동 탐지 챌린지에서 제시하는 평가 기준과 방법이 있어서 대부분의 제안된 모델들이 동일한 평가 방법과 기준으로 모델들이 평가되었다.

공간적 객체 탐지 모델의 영향을 많이 받아 2단계 시간적 행동 탐지 모델들이 제안되었고, 시각 정보를 이용해 프로포절을 생성하는 스코어 기반의 행동 국지화 모델들이 등장하였는데, 특히 2차원 스코어 맵 기반의 행동 국지화 모델들이 현재 대체로 높은 성능을 기록하고 있다.

그러나 이미 정체기에 온 객체 탐지 기술 또는 UCF-101[1] 데이터셋에 대한 행동 인식 기술과는 대조적으로, 행동 탐지 및 국지화 기술은 THUMOS-14에 대해서 아직 tIoU=0.5에서 mAP가 50%를 넘기지 못해서 실세계 응용에서 널리 사용되기 위해서는 더 큰 발전이 필요한 상태이다. 앞으로 객체 탐지나 행동 인식에서 성공적으로 사용된 다양한 방법들을 도입해서 더 높은 성능을 달성하는 행동 탐지 및 국지화 모델들이 제안될 것으로 예상된다.

### 용어해설

**행동 탐지** 무편집 비디오에서 행동이 발생한 시간 구간을 추정하고, 그 구간이 나타내는 행동 클래스를 인식하는 기술

**행동 국지화** 무편집 비디오에서 행동이 발생한 시간 구간을 추정하는 기술

**행동 인식** 하나의 행동만 포함되도록 잘 편집된 비디오의 행동 클래스를 분류하는 기술

**NMS** 임계치 이상으로 중첩된 시간 구간들 중에서 신뢰도가 가장 높은 시간 구간 하나를 제외하고 나머지 시간 구간을 제거하는 알고리즘

### 약어 정리

AR	Average Recall
AUC	Area Under the Curve
CNN	Convolutional Neural Network
GCN	Graph Convolutional Network
GT	Ground-truth
LSTM	Long Short-Term Memory Model
mAP	mean Average Precision
NMS	Non-Maximum Suppression
RNN	Recurrent Neural Network
TAG	Temporal Actionness Grouping
tIoU	temporal Intersection Over Union

### 참고문헌

- [1] K. Soomro et al., "UCF101: A dataset of 101 human actions classes from videos in the wild," CoRR, abs/1212.0402, 2012.
- [2] X. Peng et al., "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," CoRR, abs/1405.4506, 2014.
- [3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," NIPS, 2014, pp. 568-576.
- [4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," CVPR, 2017, pp. 4724-4733.

- [5] S. Asghari-Esfeden et al., "Dynamic motion representation for human action recognition," WACV, 2020, pp. 557-566.
- [6] L. Wang et al., "Action recognition and detection by combining motion and appearance features," ECCV THUMOS Workshop, 2014.
- [7] D. Oneasta et al., "The LEAR submission at THUMOS 2014," ECCV THUMOS Workshop, 2014.
- [8] S. Karaman et al., "Fast saliency-based pooling of fisher encoded dense trajectories," ECCV THUMOS Workshop, 2014.
- [9] Y.-G. Jiang et al., "Challenge: Action recognition with a large number of classes," ECCV THUMOS Workshop, <http://csrc.ucf.edu/THUMOS14/>, 2014.
- [10] A. Montes et al., "Temporal activity detection in untrimmed videos with recurrent neural networks," the 1st NIPS Workshop on Large Scale Computer Vision Systems, 2016.
- [11] S. Ma et al., "Learning activity progression in LSTMs for activity detection and early detection," CVPR, 2016, pp. 1942-1950.
- [12] B. Singh et al., "A multi-Stream bi-directional recurrent neural network for fine-grained action detection," CVPR, 2016, pp. 1961-1970.
- [13] R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," CVPR, 2014, pp. 580-587.
- [14] R. Girshick, "Fast R-CNN," ICCV, 2015, pp. 1440-1448.
- [15] S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," NIPS 2015.
- [16] Z. Shou et al., "Temporal action localization in untrimmed videos via multi-stage CNNs," CVPR 2016, pp. 1049-1058.
- [17] D. Tran et al., "Learning spatiotemporal features with 3D convolutional networks," ICCV, 2015, pp. 4489-4497.
- [18] Y. Zhao et al., "Temporal action detection with structured segment networks," ICCV, 2017, pp. 2914-2923.
- [19] K. He et al., "Spatial pyramid pooling in deep convolutional networks for visual recognition," ECCV, 2014, pp. 346-361.
- [20] J. Gao et al., "Cascaded boundary regression for temporal action detection," BMVC, 2017.
- [21] J. Gao et al., "TURN TAP: Temporal unit regression network for temporal action proposals," ICCV, 2017, pp. 3628-3636.
- [22] H. Xu et al., "R-C3D: Region convolutional 3D network for temporal activity detection," ICCV, 2017, pp. 5783-5792.
- [23] X. Dai et al., "Temporal context network for activity localization in videos," ICCV, 2017, pp. 5793-5802.
- [24] J. Gao et al., "CTAP: complementary temporal action proposal generation," ECCV, 2018.
- [25] T. Lin et al., "BSN: Boundary sensitive network for temporal action proposal generation," ECCV, 2018.
- [26] Y. Liu et al., "Multi-granularity generator for temporal action proposal," CVPR, 2019, pp. 3604-3613.
- [27] T. Lin et al., "BMN: Boundary-matching network for temporal action proposal generation," ICCV, 2019, pp. 3889-3898.
- [28] H. Eun et al., "SRG: Snippet relatedness-based temporal action proposal generator," IEEE Trans. circuits and systems for video technology(TCSVT), Early Access, 2019.
- [29] C. Lin et al., "Fast learning of temporal action proposal via dense boundary generator," AAAI, 2020.
- [30] W. Liu et al., "SSD: Single shot multibox detector," ECCV, 2016.
- [31] T. Lin et al., "Single shot temporal action detection," MM, 2017.
- [32] D. Zhang et al., "S3D: single shot multi-span detector via fully 3D convolutional network," BMVC, 2018.
- [33] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," ICRL, 2017.
- [34] S. Yan et al., "Spatial temporal graph convolutional networks for skeleton-based action recognition," AAAI, 2018.
- [35] X. Wang and A. Gupta, "Videos as space-time region graphs," ECCV, 2018.
- [36] R. Zeng et al., "Graph Convolutional networks for temporal action localization," ICCV 2019, pp. 7094-7103.
- [37] C. Zhai et al., "Action co-localization in an untrimmed video by graph neural networks," MMM, 2020.
- [38] <http://activity-net.org/challenges/2019/challenge.html>
- [39] F.C. Heilbron et al., "ActivityNet: A large-scale video benchmark for human activity understanding," CVPR, 2015.
- [40] L. Wang et al., "Untrimmednets for weakly supervised action recognition and detection," CVPR, 2017, pp. 4325-4334.