# 프레임 간 불확실성 완화를 위한 어텐션 마스킹 기반 시간적 행동 검출 연구

강태경, 이승원, 백승호
LIG 넥스원

taekyung.kang@lignex1.com, seungwon.lee@lignex1.com, seungho.baek@lignex1.com

# Attention masking-based temporal action detection for inter-frame ambiguity mitigation

Taekyung Kang, Seungwon Lee, Seungho Baek
LIG Nex1

요 약

시간적 행동 검출(temporal action detection)은 복잡하고 가공되지 않은 비정형 비디오 내에서 다양한 사람 행동을 식별하는 도전적인 컴퓨터 비전 분야 과제이다. 기존 연구들은 주로 사전 학습된 비디오 인코더를 활용하여 스니펫(snippet) 단위의 영상 특징을 추출해 왔다. 그러나 이러한 스니펫 단위 특징은 프레임 간의 시간적 정보의 부족으로 인해 행동 발생 시점을 정확히 예측하는 데 어려움이 있다. 본 논문에서는 이러한 문제를 해결하기 위해, 프레임 간 불확실성과 모호성을 완화하는 어텐션 마스킹 기법을 제안한다. 행동 인식 어텐션 모듈을 설계하여 행동의 속성을 반영한 마스크를 생성하고 이를 통해 스니펫 단위 영상 특징을 주변 행동의 문맥 정보를 고려하여 주요 행동 영역을 더욱 두드러지게 특징을 재구성한다. 본 기법은 구성요소 제거 실험을 통해 모듈의 기여도를 검증하고, 전체 시스템의 성능 향상 효과를 입증하였다.

## Ⅰ. 서 론

시간적 행동 검출(temporal action detection, TAD)은 비디오 이해(video understanding) 분야에서 중요한 도전 과제 중 하나이다. 시간적 행동 검출 연구의 목표는 가공되지 않는 비정형(untrimmed) 비디오 내에서 행동의 시작 및 종료 시점, 그리고 행동의 클래스를 예측하는 것으로, 이는 비디오 요약, 비디오 감시, 비디오 검색 등 다양한 응용 분야로 확장될 수 있다. 정형(trimmed) 비디오를 입력으로 하여 행동 분류(classification)만을 수행하는 행동 인식 연구와는 다르게, 긴 영상에서 복잡하고 모호한 행동을 식별하기 위해 행동 분류와 동시에 시간 탐지(localization) 작업을 수행해야 한다.

최근에는 다양한 효과적인 시간적 행동 검출 기법들이 제안되었으며, 대부분은 사전 학습된 비디오 인코더를 기반으로 동작한다. 일반적으로 비정형 비디오를 스니펫 단위로 분할하여 영상 특징을 추출하고, 이를 바탕으로 행동 검출 모델이 시간적 행동 경계와 행동 클래스 예측을 수행한다. 기존 방법들이 우수한 성능을 보였음에도 불구하고, 제한적인 특징 표현력으로 인해 영상 내 내재한 의미적 정보를 충분히 활용하지 못하는 한계가 존재한다. 특히, 비디오 단위 분류 작업인 행동 인지 작업을 위해 설계된 사전 학습 인코더는 시간적 행동 검출 작업에 최적화되어 있지 않아, 스니펫 단위로 추출된 특징들이 충분한 문맥 정보를 포함하지 못하는 문제가 발생한다.

일반적으로 스니펫 단위 영상은 8 프레임에서 32 프레임 정도로 구성되는데, 이는 30fps 기준 0.27 초에서 1.07 초로 매우 짧은 시간 정보를 포함한다. 이로 인해 연속적인 프레임의 특징 정보 간의 모호성이 발생하게 되어 행동이 발생하는 행동 프레임과 행동이 발생하지 않는 배경 프레임을 명확히 구분하는 데 어려움을 초래하며, 결과적으로 행동 검출 및 분류 성능에 부정적인 영향을 미친다. 또한, 이러한 프레임 간 모호성은 분류와 시간적 행동 경계 예측 간의 불일치를 유발할 수 있다. 예를 들어, 예측된 시간적 행동 경계가 정확 하더라도, 분류 점수의 부정확성으로 인해 논 맥시멈 서프래션(non-maximum suppression, NMS[1]) 과정에서 해당하는 행동 구간이 제거되어 최종 검출 성능이 저하될 수 있다.

본 논문에서는 프레임 간 모호성 문제를 해결하기 위해 어텐션 마스킹 기법을 제안한다. 제안하는 방법은 기존 방법에 비해 유연하고 강인한 어텐션 마스크를 학습하여, 기존 스니펫 단위 특징의 표현력을 향상하고 최종적으로 시간적 행동 검출 성능을 개선한다.

### Ⅱ. 1. 문제 정의 및 특징 추출

본 논문에서는 시각적 행동 검출을 위한 프레임워크를 제안하며, 이는 행동 인식 어텐션 모듈과 예측 헤드로 구성된다. 가공되지 않은 비정형 비디오 입력으로부터 행동의 시작과 끝 시점, 그리고 해당 행동의 클래스를 예측하는 것이 본 과제의 목표이다. 이를 위해 먼저 사전 학습된 비디오 인코더(I3D[2])를 활용하여, 일정 길이(예: 16 프레임)로 분할된 스니펫 단위 영상에서 영상 특징 $x \in \mathbb{R}^{T \times C}$ 을 추출한다. 여기서 $T$ 는 시간 차원, $C$ 는 채널 수를 의미한다.

### Ⅱ. 2. 마스크 표현 학습

본 연구에서 행동과 배경 간의 특성을 구분할 수 있는 어텐션 마스크를 학습하기 위해 **그림 1** 과 같은 구조로 설계하였다. 우선, 스니펫 단위 특징을 행동

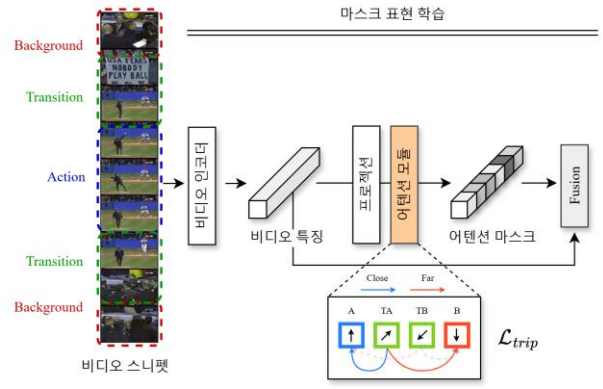


그림 1 어텐션 마스크 학습 예시

영역(positive)과 배경 영역(negative)으로 구분한다. 이후 행동 영역을 강조하고 배경 영역과의 임베딩 거리 차이를 유지하는 방식으로 어텐션 마스크를 학습한 뒤, 본래 특징과 융합되어 행동 영역을 더욱 두드러지도록 재구성한다. 비디오 내 행동 특징과 배경 특징은 상호 독립적인 특성을 가져야 하며, 어텐션 마스크는 이 두 영역을 명확히 표현할 수 있어야 한다. 이를 위해, 모델이 안정적으로 학습을 할 수 있도록 행동과 배경 사이의 경계 부분을 세분화시켜 행동 전이(transition action, TA)와 배경 전이(transition background, TB) 영역을 추가로 정의하였다. TA 는 임베딩 공간에서 행동 영역에 가깝고 배경에서 멀어지도록, TB 는 배경에 가깝고 행동에서 멀어지도록 설계된 손실 함수를 적용함으로써, 연속된 프레임 간 모호성을 완화하고 유연한 어텐션 마스크 생성을 유도한다.

이를 위해 전체 영상 특징 $x$ 를 1D 컨볼루션 기반 인코더 $f(\cdot)$ 를 통해 임베딩 공간으로 투영하고, 해당 임베딩을 다음 네가지 구성요소로 분할한다: 행동($y^a$), 배경($y^b$), 행동 전이($y^{ta}$), 배경 전이($y^{tb}$).

최종적으로 학습은 삼중항 손실함수(triplet loss)를 기반으로 진행되며, 구성 요소 간 임베딩 거리 차이를 통해 행동과 배경, 전이 영역 간의 구분을 더 명확히 할 수 있도록 마스크를 학습한다. 삼중항 손실 함수는 다음과 같이 구성된다.

$$\mathcal{L}_{trip}(a,p,n) = [\|a-p\|_2^2 - \|a-n\|_2^2 + \alpha]_+,$$

여기서 $\alpha$는 positive 샘플과 negative 샘플 간의 최소 거리 차이를 유지하기 위한 마진(margin) 값이다.

각 구성 요소 간의 특성을 학습하기 위해, 전체 표현 손실 함수 $\mathcal{L}_{rep}$는 다음과 같이 구성된다.

$$\mathcal{L}_{rep} = \mathcal{L}_{trip}(y^{ta}, y_a, y^b) + \mathcal{L}_{trip}(y^{tb}, y^b, y^a) + \lambda_m \cdot \mathcal{L}_{trip}(y^m, y^a, y^b),$$

여기서 $y = f(x)$ , $y^m$ 는 어텐션 마스크 특징, $\lambda_m$ 은 어텐션 마스크에 대한 가중치이다. 해당 손실함수 기반 학습을 통해 비디오 특징 내 행동 영역에서의 표현력을 향상시키고, 연속된 프레임 간 모호성을 효과적으로 완화한다.

## Ⅱ. 3. 데이터셋 및 실험

제안하는 어텐션 마스크 효과를 검증하기 위해, THUMOS14[3] 벤치마크 데이터셋을 기반으로 다양한 어텐션 구성 요소에 대한 구성요소 제거 실험(ablation study)을 수행하였다. 행동 검출 성능 평가는 클래스별 평균 정밀도(mean Average Precision, mAP)를 기준으로 수행하였다. 이는 각 행동 클래스에 대한 평균 정밀도 값을 산출한 후, 그 평균을 계산하는 방식이다. 평가

표 1 구성요소 제거 실험

| 구분 | mAP@tIOU (%) | | | | | |
|---|---|---|---|---|---|---|
| | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 평균 |
| Baseline | 73.9 | 68.1 | 61.3 | 48.2 | 32.1 | 56.7 |
| + A, B | 76.7 | 73.1 | **66.8** | 57.2 | **42.7** | 63.3 |
| + A, B, T | 76.8 | 72.8 | 65.9 | 56.9 | 41.6 | 62.1 |
| + A, B, TA, TB | **77.5** | **73.7** | 66.3 | **57.3** | 41.9 | **63.7** |

기준으로는 시간적 교집합 비율(temporal Intersection over Union, tIoU) 임계 값을 [0.3:0.1:0.7] 구간으로 설정하여 측정하였다.

**표 1** 의 두 번째 행(+ A, B)은 행동과 배경 영역만으로 구성한 경우이며, 세 번째 행에서는 전이 영역(+ A, B, T)을 추가하였다. 그러나 이 경우, 전이 영역에 대한 학습이 불안정하여 성능이 오히려 저하되는 결과를 보였다. 반면, 네 번째 행(+ A, B, TA, TB)에서는 세부적으로 행동 전이와 배경 전이로 구성함으로써 베이스라인 기준 mAP@Avg. 기준 +7%p 성능 향상을 이루었다.

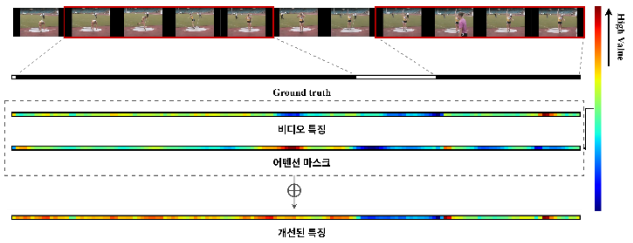


그림 2 어텐션 마스킹 과정 시각화

**그림 2** 는 어텐션 마스킹 과정을 시각화한 결과이다. 기존 비디오 특징은 사람의 행동이 발생하는 구간에서의 특징이 뚜렷하지 않은 것을 확인할 수 있다. 반면, 어텐션 마스크를 적용한 특징은 행동 영역 주변에서 더욱 뚜렷한 특징 정보를 나타내며, 비디오 특징이 시간적 의미 정보를 효과적으로 활용할 수 있도록 개선되었음을 보였다.

## Ⅲ. 결론

본 논문에서는 시간적 행동 검출 과정에서 발생하는 프레임 간 모호성을 완화하기 위해 행동 인식 기반 어텐션 마스킹 기법을 제안하였다. 제안하는 방법은 행동과 배경 영역을 구분하여 특징을 재구성하고, 행동 전이 영역을 고려한 손실함수를 통해 보다 유연하고 견고한 어텐션 마스크를 학습한다. 이를 통해 행동 발생 영역의 표현력을 강화하고, 행동 검출의 정확도를 향상시킬 수 있음을 실험을 통해 확인하였다. 향후 다양한 시각적 행동 패턴을 포괄하는 확장 연구를 통해 본 방법의 범용성을 더욱 강화할 예정이다.

## 참 고 문 헌

[1] Bodla, Navaneeth, et al. "Soft-NMS--improving object detection with one line of code." in ICCV, 2017.

[2] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." in CVPR, 2017.

[3] Idrees, Haroon, et al. "The thumos challenge on action recognition for videos “in the wild”." Computer Vision and Image Understanding 155, 1-23, 2017.