졸업논문 소개

글로벌-로컬 뷰 정렬과 배경 편향 제거를 통한 대규모 도메인 격차를 가진 비지도 비디오 도메인 적응

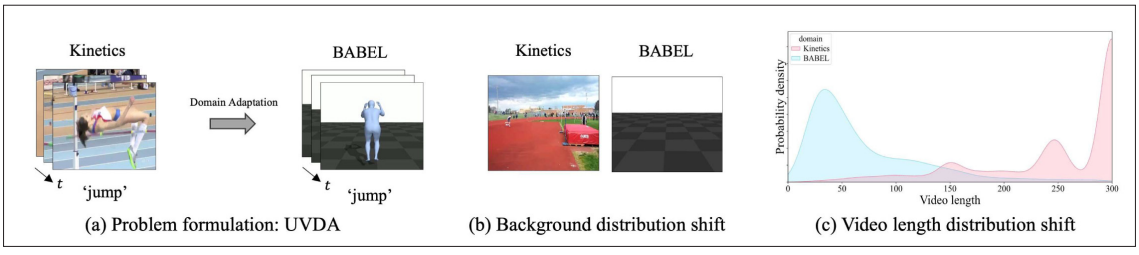
이효건 / 경희대학교 Vision & Learning Lab.

인간 행동 인식은 활발하게 연구되고 있는 컴퓨터 비전의 하위 분야이다. 비디오 영상 감시 체계, 스포츠 분석, 헬스케어, 자율주행 등 다양한 분야에서 쓰인다. 따라서, 더 강건하고 성능이 좋은 모델은 해당 분야에 큰 영향을 끼칠 수 있다.

이전 연구에서는 합성곱 신경망(CNNs; Convolutional Neural Networks)이나 트랜스포머를 이용한 방법 등의 인공 신경망을 이용한 방법들이 제시되었다. 이러한 인공 신경망 방법들은 대규모의 데이터셋을 이용한 최적화 기반 학습 과정을 거치며, 이때 사용하는 데이터셋은 각 비디오마다 라벨(label)이 기록되어 있어야 한다. 하지만, 비디오는 이미지보다 복잡하여 라벨링 시 시간이 많이 들며, 이는 비용 문제와 직결된다.

최근 연구에서는 라벨링 문제를 풀기 위한 방법 중 하나로 비지도 도메인 적응(UDA; Unsupervised Domain Adaptation)이 제시되었다. UDA는 라벨이 있는 원천(source) 데이터셋을 이용하여 라벨이 없는 목적(target) 데이터셋에서 좋은 성능을 보이는 것을 목표로 한다. 최근 연구는 이를 비디오에 적용한 비지도 비디오 도메인 적응(UVDA; Unsupervised Video Domain Adaptation) 방법을 통해 뛰어난 결과를 보여주었다.

하지만 기존의 UVDA 방법론들은 도메인 격차가 적은 데이터셋에서만 실험이 이루어져 좋은 수치적 성능을 보여주는 데에도 불구하고 실제 성능은 그렇게 좋지 못하다.



<그림 1> 키네틱스 → 바벨의 개요

일례로 UCF-HMDB 데이터셋은 UCF101과 HMDB51의 별개의 두 행동 인식 데이터셋을 재구성하여 UVDA 세팅으로 만든 데이터셋인데, 비디오 개수가 3,209개 밖에 없어 매우 작고, 비슷한 시기에 비슷한 소스로부터 수집했기에 영상들의 모습도 굉장히 비슷하다. 실제로, UCF-HMDB의 목적 데이터셋에서 학습한 단순 행동 인식 모델과 원천 데이터셋에서 학습한 모델의 성능 차이는 단 11.4%p밖에 나지 않는다. 이 둘은 각각 UVDA 모델이 가질 수 있는 성능의 상한/하한으로 볼 수 있는데, 차이가 작다는 것은 문제가 쉬움을 뜻한다. 반면, 실제 문제는 날씨, 조도, 시야 차이 등의 해당 데이터셋에서 다루지 않은 다양한 어려운 점들이 존재한다.

본 연구는 해당 문제들을 골고루 모델링하는 키네틱스→바벨 데이터셋을 제시한다. 〈그림 1〉에서와 같이, 키네틱스→바벨은 두 도메인 간 큰 배경 차이와 비디오 길이 차이를 보여준다. 키네틱스→바벨은 총 18,694편의 비디오 영상으로 구성되며, 기존 데이터셋 대비 상당히 큰 규모이다. 〈표 1〉에 이러한 차이를 수치적으로 제시한다.

<표 1> UVDA 데이터셋의 통계치

데이터셋	클래스 수	비디오 수	배경 차이	길이 차이	정확도 차이
UCF-HMDB	12	3,209	0.17	90.9	11.4
EPIC-KITCHENS	8	* 6,729	0.11	62.7	26.2
Mixamo→Kinetics	**14**	**36,195**	0.24	66.7	**†68.1**
Kinetics→BABEL	12	18,946	**0.31**	**182.1**	65.0

* 6 개의 세팅에 대한 평균이다.
† 해당 논문에 기재된 값이다.

또한, 본 연구는 대규모 도메인 격차를 가진 키네틱스→바벨을 풀기 위한 방법으로 글로벌-로컬 뷰 정렬과 배경 편향 제거 방법을 제시한다.

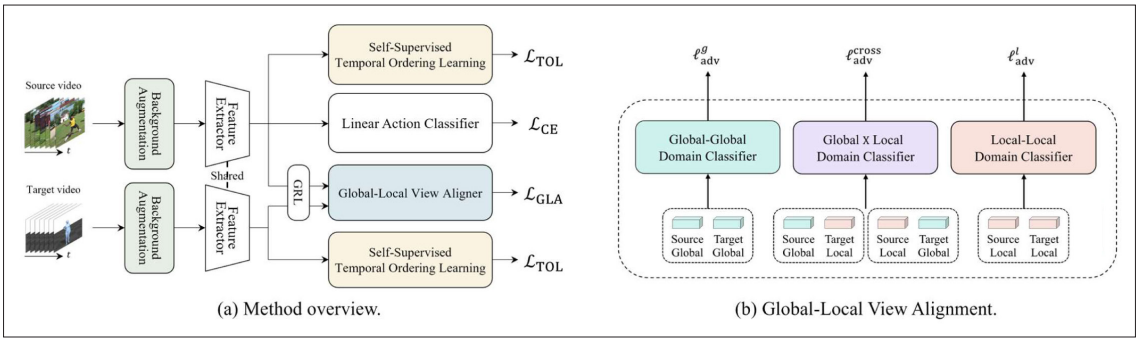
본 연구는 UVDA에서 마주하는 실제적 상황을 더 잘 모델링하기 위해 키네틱스→바벨 데이터셋을 제시한다. 키네틱스→바벨은 원천 도메인으로 키네틱스, 목적 도메인으로 바벨 데이터셋을 사용한다. 해당 데이터셋은 키네틱스와 바벨을 재구성하여 만들어졌다. 키네틱스→바벨은 총 12개의 범주를 갖고 있으며 이는 다음과 같다: jump, run, throw, kick, bend, dance, clean something, squat, punch, crawl, clap, pick up. 각 훈련, 평가 데이터셋에 포함된 영상 개수는 키네틱스가 14,881개, 650개, 바벨이 2,943개, 452개이다.

제안된 데이터셋의 키네틱스는 현실적 영상이고, 바벨은 모션 캡쳐 기반으로 합성된 영상이다. 합성 영상은 라벨링이 쉽기 때문에 이를 채택하는 것은 자연스러운 과정이다. 이러한 시도는 이전에도 있었으며, 현실→합성, 합성→현실의 양방향 모두 상당히 어려운 문제임을 시사한다. 본 연구는 전자인 키네틱스→바벨에 중점을 두었다.

한편, 키네틱스→바벨의 두 도메인은 상당한 격차를 보여준다. 본 연구는 도메인 간 격차를 형상 격차와 시간 길이 격차로 나누었다. 키네틱스→바벨은 두 격차가 모두 크다. 바벨의 경우 3D 그리드에 모션 캡쳐 데이터를 재구성했기에 배경이 아예 없다. 하지만 키네틱스는 풍부한 현실적 배경들을 갖고 있다. 게다가, 키네틱스와 바벨은 상당히 다른 시간 길이 분포를 보인다. 이로서 해당 데이터셋은 높은 난도를 갖게 된다.

특히, 키네틱스→바벨은 UCF-HMDB, EPIC-KITCHENS와 같은 UVDA 데이터셋들보다 상당한 도메인 격차를 보인다. 형상 격차를 수치화하기 위하여, 본 연구는 두 도메인 간 평균 최소 배경 피처 거리인 배경 격차를 정의한다. 이는 한 도메인의 배경 피처마다 다른 도메인에서 제일 가까운 배경 피처를 골라 잰 코사인 거리의 평균이다. 배경 피처는 Places365로 학습된 ResNet-50 모델을 사용하여 추출했다.

또한, 키네틱스→바벨의 시간 길이 차이 또한 해당 데이터셋들보다 크다. 이를 수치화하는 데에는 EMD(the earthmover's distance) 거리를 사용했다. EMD 거리는 한 확률분포로부터 다른 확률분포로 변환될 때의 최소한



<그림 2> 제시된 방법론의 개요

의 비용을 나타내며, 분포 간 차이를 직관적으로 나타낸다. 본 연구는 두 도메인의 영상 길이 분포를 확률분포 p, q로 나타내었고, 이 둘의 시간 길이 격차를 EMD(p,q)로 정의했다.

각 데이터셋의 도메인 격차를 비롯한 통계적 비교는 〈표 1〉에서 확인할 수 있다. 키네틱스→바벨은 상당히 큰 데이터셋이며, 배경과 시간 길이 모두 가장 큰 격차를 보인다.

UVDA는 라벨이 있는 원천 데이터셋으로부터 라벨이 없는 목적 데이터셋으로의 학습을 목표로 한다. 이 둘은 라벨 공간을 공유한다. 비디오 환경에서 단순히 원천 데이터셋에서 학습하고 목적 데이터셋에서 평가하면 심각한 배경, 시간 길이 편향이 발생하게 된다. 이를 해결하기 위해 본 연구는 i) 글로벌-로컬 뷰 정렬, ii) 배경 증대, iii) 자기지도 순서 학습의 세 방법을 제시한다. 해당 방법들의 개요는 〈그림 2〉에서 확인할 수 있다.

i) 같은 행동 클래스여도 키네틱스와 바벨에서 그 시간적 길이가 다르다. 가령 jump라는 클래스는 키네틱스에서 10초이지만, 바벨에서는 1초 정도밖에 되지 않는다. 따라서, 시간 길이에 대한 편향을 제거할 필요가 있다.

한편, 균등 샘플링은 비디오를 균등하게 분할하여 분할된 클립마다 프레임을 한 장씩 뽑는다. 반면, 고밀도 샘플링은 비디오 길이에 상관없이 일정한 간격으로 프레임을 뽑는다. 본 연구는 두 샘플링 기법이 각각 글로벌, 로컬 특징을 표현할 수 있다는 점에서 착안하여 글로벌-로컬 뷰 정렬 방법을 제시한다.

M개의 글로벌 피처와 N개의 로컬 피처 각각의 평균 피처인 조합 피처를 얻을 수 있다. 이들을 글로벌, 로컬, 크로스 3개의 도메인 분류기에 입력으로 넣어 적대적 학습하는 방식으로 도메인 피처를 정렬한다. 다양한 시간 길이 조합을 사용하기 때문에 시간 길이 편향을 제거할 수 있다.

ii) 〈그림 1〉(a)에서와 같이 키네틱스→바벨은 배경을 통한 힌트를 사용하는 것이 불가능하다. 따라서 배경 편향을 제거할 수 있는 방법이 필요하다. 동영상의 시간 축에 중간값 필터를 적용하면 배경 이미지를 쉽게 얻을 수 있다. 사전에 모든 영상에 대한 배경을 다 추출하여 배경 DB를 구축한다. 학습 시에는 배경 DB로부터 무작위로 한 장을 추출하여 동영상 입력에 blend하여 증대시킨다. 단순한 방법이지만, 배경에 관한 편향을 상당히 제거할 수 있다.

배경 편향 제거를 위한 두 번째 방법으로 순서 학습을 생각할 수 있다. 동영상의 무작위 N개의 클립을 뽑아 순서를 맞추게 한다. 순서를 잘 맞히려면 배경이 아닌 사람과 같은 움직이는 부분에 집중해야 하기 때문에 배경 편향

<표 2> 각 요소에 대한 에블레이션

방법		MCA
편향제거	글로벌-로컬 뷰 정렬	K→B
		26.4 ± 2.4
✓		36.7 ± 3.6
	✓	26.9 ± 3.2
✓	✓	**37.7** ± 2.5

<표 3> 최신 모델들과의 비교

방법	사용한 프레임 수	K→B
Source-only	3 x 8 = 24	11.7 ± 0.7
DANN	3 x 8 = 24	29.3 ± 1.5
CoMix	16 x 8 = 128	21.4 ± 0.3
CO2A	4 x 16 = 64	24.7 ± 0.8
Ours	3 x 8 = 24	**33.7** ± 1.8
Target-only	3 x 8 = 24	76.7 ± 2.1

제거 효과를 기대할 수 있다.

각 요소에 대한 기여도를 알아보기 위해 에블레이션 및 최신 모델들과 비교 실험을 진행하여 〈표 2〉, 〈표 3〉에 제시했다. 글로벌-로컬 뷰 정렬과 배경 편향 제거 방법은 각각의 효과도 있었지만 둘을 같이 사용했을 때 더 큰 효과를 보았다. 또한, 다른 최신 모델들과의 비교에서 샘플링한 클립 개수, 클립 당 프레임 수를 비교했을 때, CO2A, CoMix 대비 훨씬 적은 프레임을 사용했음에도 더 높은 성능을 보였다.

이 효 건

- 2021년 8월 : 경희대학교 전자공학과 학사
- 2024년 2월 : 경희대학교 컴퓨터공학과 석사
- 2024년 3월 ~ : 경희대학교 산학협력단 연구원
- 주관심분야 : 비디오 이해, LLMs