야구 중계 영상 내 행동 인식을 위한 모델 및 데이터셋 비교 분석 연구

박정현[1], 김수경[1], 김대한[1], 이승현[1], 신윤호[2], 이웅희[1]*
[1]한성대학교, [2]엘지유플러스

[1]{ shshjhjh4455, 2171175, kkkhaeun, puppy1012, whlee}@hansung.ac.kr, [2]yoon731@gmail.com

Comparative Analysis of Models and Datasets for Action Recognition in Baseball Broadcast Videos

Jeonghyeon Park[1], Su-Gyeong Kim[1], Daehan Kim[1], Seungheyon Lee[1], Yoonho Shin[2], Woonghee Lee[1]*
[1]Hansung University, [2]LG Uplus

요 약

본 연구에서는 다양한 야구 중계 영상에서 투수 및 타자 등과 같은 인물을 구별할 수 있는 행동 인식 모델을 탐색 및 분석한다. 그 다음, 현재 오픈 소스로 공유되어 있는 여러 야구 영상과 관련한 데이터셋을 비교하여 야구 경기 영상 내에서 행동을 인식하고 이해하여 해당 인물이 어떤 포지션에 있는지 구별하는데 좋은 성능을 내는 데 가장 적합한 데이터셋을 찾아본다. 이를 통해 행동 인식 모델을 이용하여 야구 경기 영상 내 투수와 타자 등 다양한 인물을 인식하고, 경기의 하이라이트 장면을 추출하기 위한 적합한 모델을 구현하는데 기여를 할 수 있을 것으로 기대된다.

Ⅰ. 서 론

HAR(Human Action Recognition)은 이미지나 영상 등의 데이터에서 다양한 방법으로 사람의 행동 정보를 수집 및 처리하여 수행중인 행동을 식별하는 기술이다. HAR 은 인간의 행동을 인식하고 이해한다는 점에서 여러 분야에 사용되고 있다. 특히 스포츠 분야에서 HAR 은 선수 감지, 움직임 추적, 수행된 동작 인식, 다양한 동작 비교 등 여러 역할을 수행한다[1]. 여러 스포츠들 중 야구는 대중에게 잘 알려져 있으며, 많은 팬을 보유하고 있다. 또한, 경기 영상의 하이라이트 영상을 팬 뿐만 아니라 많은 대중들도 관심있게 시청한다. 그래서 본 연구는 야구 중계 영상에서 하이라이트 장면을 인식 및 이해하여 5~10 초 정도의 하이라이트 영상을 추출하는 기술을 구현하는데 기여하는 것을 목표로 한다. 이를 위해 해당 목표에서 이뤄야 할 가장 기초적인 부분인 투수, 타자 및 포수처럼 경기 내 등장하는 인물을 식별하는데 가장 적합한 모델을 찾아본다. 그리고 해당 모델을 학습시키는데 적합한 데이터셋을 비교 분석하여, 최적의 모델을 구현하기 위한 방향을 모색한다.

Ⅱ. 본론

본 연구에서는 크게 RGB 기반의 모델들 중 가장 좋은 성능을 내는 모델 한 개와 Skeleton 기반의 모델들 중 가장 좋은 성능을 내는 모델 한 개를 비교 분석하여, 야구 영상에서 행동을 인식·이해하여 영상에 등장하는 인물을 구별하는데 가장 적합한 모델을 찾아보았다. 비교를 위해 mmaction2 에서 제공하는 pre-trained model 을 이용해 두 모델을 테스트해본다. mmaction2 는 홍콩 중문 대학교 멀티미디어 연구소에서 개발한 비디오 내 행동 인식에 특히 초점을 맞춘 open source tool box 이며, 행동 인식·이해 연구를 위해 설계되었다. 행동 인식에 대한 개발 및 연구를 촉진하기 위해 사전 구축된 다양한 모델, 데이터셋 및 평가 도구를 함께 제공하고 있으며, 비디오 내 행동 인식에 관심이 있는 여러 사용자가 이용하고 있다.

mmaction2 에서 제공하는 여러 행동 인식 모델을 비교하기 위해 kinetics-400 데이터셋에서 임의로 10 개를 뽑아 평균 정확도를 계산해보았다. 그 결과 TSN(Temporal Segment Network)[2] 모델이 가장 높은 정확도를 나타냈다. TSN 모델은 long-range temporal structure modeling 아이디어를 기반으로 하며, 비디오 기반 행동 인식에 새로운 프레임워크를 제시한다. 이 모델은 전체 비디오를 사용해 행동을 인식하는데, 효율적이고 효과적인 학습을 가능하게 하도록 sparse temporal sampling 전략과 video-level supervision (비디오 수준의 감독)을 포함한다[2]. long-range temporal structure 를 활용할 수 있다는 점이 큰 특징이며, 그림 1 은 학습된 ConvNet 모델을 시각화한 모습으로서 TSN 의 효율성과 모범 사례를 보여준다.

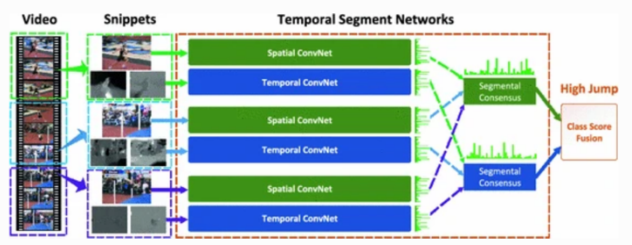

그림 1. Temporal Segment Networks 구조[2]

이러한 RGB 기반의 행동 인식 모델은 야구 중계 영상마다 구도나 형식이 다르면 정확도나 성능이 저하될 수도 있다는 점에서 문제가 있다고 판단됐다. 그리고 야구 경기에서 단순히 공을 던지는 행위는 투수 뿐만 아니라 수비수도 동일한 행동을 보여주며, 단지 공을 던지는 폼의 차이가 존재할 뿐이다. 따라서 동작을 더 정확히 인식할 수 있는 모델을 찾으려 하였고, 그 결과 skeleton 기반 행동 인식 모델을 사용하기로 결정하였다.

mmaction2 에서 제공하는 skeleton 기반 행동 인식 모델 중 성능이 좋으면서 앞서 제기된 문제를 해결하기에 가장 적합한 모델은 PoseC3D 라 판단했다. PoseC3D 모델은 그래프 시퀀스 대신 3D 히트맵 스택을 사용한다. GCN(Graph Convolutional Networks)[3] 기반 방법과 비교해, PoseC3D 는 시공간적 특징들을 학습하는데 더욱 효과적이고, 포즈 추정 노이즈에 더욱 견고하며 교차 데이터셋에서 더 잘 일반화된다. 그리고 여러 사람이 참여하는 시나리오를 추가 계산 비용없이

효율적으로 처리하고, 다른 모달리티와도 잘 통합되어 까다로운 데이터셋에서 지속적으로 강력한 성능을 달성한다[3].

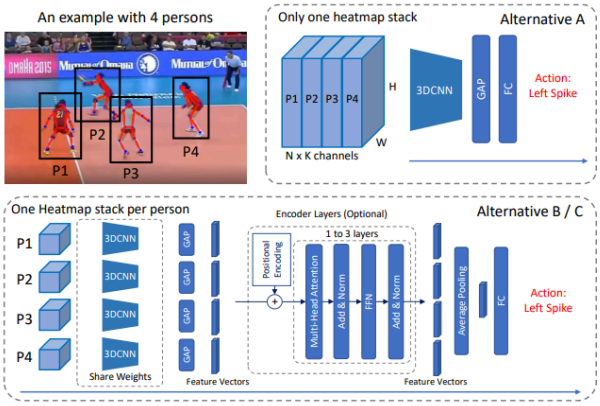

그림 2. PoseC3D Architecture[3]

우리는 PoseC3D 모델이 (1) 여러 사람이 참여하는 시나리오를 효율적으로 처리한다는 점에서 여러 선수가 등장하는 야구 영상을 효율적으로 처리할 수 있다고 보았고, (2) 까다로운 데이터셋에서 지속적으로 강력한 성능을 달성한다는 점에서 방송사별 다른 형식의 야구 영상에서도 강력한 성능을 보일 것이라 판단했다. 따라서 PoseC3D 모델이 야구 중계 영상에서 행동 인식·이해를 통해 인물을 구별하며, 방송사별 다른 형태의 영상에도 잘 적응할 수 있는 모델이라 판단했다.

PoseC3D 를 야구 경기 영상에서 행동 인식·이해를 통해 선수를 구별하는데 좋은 성능을 보이도록 학습시키기 위해, 우리는 야구와 관련된 데이터셋을 찾아보았다. 우리가 찾아본 데이터셋 중 오픈소스로 공유되어 있는 야구 경기 데이터셋은 MLB-YouTube, BBDB, kinetics-(400/600/700), 그리고 AI-Hub 에서 제공하는 야구 스포츠 영상 데이터 등이 있었다.

MLB-YouTube 데이터셋은 No Activity, Ball, Strike, Swing, Hit, Foul, In Play, Bunt, Hit by Pitch 로 총 9 개의 클래스로 구성되어 있다[4]. 해당 데이터셋은 1 시간 30 분이 넘는 긴 영상에서 각 레이블에 맞는 부분을 1~4 초 길이의 영상으로 클립을 생성할 수 있도록 깃허브에 코드를 제공하고 있다. 하지만 개발 환경의 차이 및 각종 오류로 인한 문제로 다른 데이터셋을 찾게 되었다.

BBDB 데이터셋은 네이버 야구 중계 영상에 레이블 된 데이터이다. BBDB 데이터셋은 시각적으로 유사한 세그먼트에 다른 레이블이 붙어 있는 등 여러 어려운 요소들을 포함하고 있다[5]. 그래서 인식, localization, 텍스트-비디오 정렬, 하이라이트 생성 등 다양한 비디오 이해 작업에 적합하다고 판단했다. 또한, 한국 야구 경기 영상이라는 점에서 우리가 원하는 작업에 유용할 것으로 기대했다. 하지만 해당 데이터셋이 만들어진 시점의 링크와 현재의 영상 링크가 달라졌다는 점과, 동일한 제목을 검색하여 영상을 찾았지만 영상의 길이 또한 차이가 있어 사용하기 어려울 것으로 판단됐다.

kinetics-(400/600/700)은 레이블이 야구를 타깃으로 한 데이터셋은 아니지만, 행동 인식 분야에서 유명한 데이터셋이며 데이터셋 내 야구 관련 영상 및 레이블이 존재한다는 점에서 사용을 하게 되었다. 하지만 야구 관련 레이블이 catching or throwing baseball 과 같이 구분해야 할 동작을 하나로 묶어 놓았다는 점에서 문제가 발생하였다. 그래서 수동으로 동영상 크롭 및 정답 레이블 txt 파일을 만들어 모델을 돌려보았지만, 모델의 성능이 낮아졌다. 그래서 우리가 원하는 작업에 효율적이고 긍정적인 영향을 주지 못할 것이라 판단되었다.

AI-Hub 에서 제공하는 야구 스포츠 영상 데이터셋은 각 관절의 위치를 수동으로 어노테이션 작업을 수행했다는 점에서 높은 정확도를 가질 것이라 판단하여 사용하게 되었다. 그러나 mmaction2 에서 제공하는 PoseC3D 모델과 해당 데이터셋의 키포인트의 포맷이 일치하지 않았다. 그래서 최종적으로 수동으로 어노테이션된 값을 사용하지 않고, mmaction2 에서 제공하는 비디오에서 어노테이션 값을 추출하는 파이썬 파일인 ntu_pose_extraction.py 를 이용하기로 결정했다. 이렇게 하면 새로운 동영상에 대해서도 키포인트 값을 추출할 수 있어 유용할 것이라 판단했다. 해당 파이썬 파일은 동영상을 받아, 그 동영상에 대한 pkl 파일만 생성한다. 그렇기에 해당 파일을 우리가 받은 데이터셋 내 이미지 파일을 입력으로 받아, 최종 모델의 입력으로 넣을 하나의 pkl 파일을 생성하도록 수정하면 우리가 가진 데이터셋을 이용해 모델을 학습시킬 수 있을 것이라 기대된다.

Ⅲ. 결론

본 연구는 야구 중계 영상에서 행동 인식·이해를 통해 영상에 등장하는 투수, 타자 등 각 포지션에 위치한 선수를 구별하고, 방송사에 따라 서로 다른 구도와 형식을 보이는 영상에서도 지속적으로 좋은 성능을 내는 행동 인식 모델을 선정하기 위해 RGB 기반의 행동 인식 모델과 skeleton 기반의 행동 인식 모델을 각각 비교해보았다. 그 결과, 해당 목표에 PoseC3D 모델이 가장 적합하다고 판단했다. 그리고 야구 영상을 통해 모델을 학습시키기 위해 적절한 데이터셋을 비교 분석해보았다. 본 연구는 야구 경기 영상 내 행동 인식을 통해 인물을 구별하고 하이라이트 영상을 추출하는 모델을 구현하는 향후 목표에 기여할 수 있을 것이라 기대된다.

ACKNOWLEDGMENT

본 연구는 한성대학교 학술연구비 지원과제임.

참 고 문 헌

[1] Kristina Host, Marina Ivasic-Kos. “An overview of Human Action Recognition in sports based on Computer vision”

[2] Limin Wang, Yuanjun Xiong, et al. “Temporal Segment Networks: Towards Good Practices for Deep Action Recognition”

[3] Haodong Duan, Yue Zhao, et al. “Revisiting Skeleton-based Action Recognition”

[4] AJ Piergiovanni, Michael S. Ryoo. “Fine-grained Activity Recognition in Baseball Videos”

[5] Minho Shim, Young Hwi Kim, et al. “Teaching Machines to Understand Baseball Games: Large-Scale Baseball Video Database for Multiple Video Understanding Tasks”