전경 분리를 활용한 인간 행동 인식 분석 실험

박영진*, 조희섭

대구경북과학기술원 ABB연구부

*yjpark@dgist.ac.kr, mozart73@dgist.ac.kr

An Experimental Study on Human Action Recognition Using Foreground Segmentation

Young-Jin Park*, Hui-Sup Cho

Division of ABB, DGIST

요 약

본 연구는 HMDB51 데이터셋에서 선별한 11개 행동 클래스를 대상으로, 전경 분리가 3D 행동 인식 모델 성능에 미치는 영향을 분석하였다. 선정된 클래스는 drink, eat, smoke, talk, pick, smile, laugh, clap, pour, wave, shake_hands로, 일상적 제스처와 상호작용을 포함한다. 원본 데이터셋과 Detectron2 기반 전경 분리 데이터셋을 각각 SlowFast 모델에 학습 및 평가한 결과, 전경 분리의 적용은 전반적으로 성능 향상으로 이어지지 않았으며, 오히려 배경 단서 등의 소실로 성능 저하가 관찰되었다. 이러한 결과는 전경 분리가 보편적인 성능 개선 기법이 아님을 보여주며, 행동 인식에서 전경 및 배경 단서가 상호보완적으로 작용함을 규명하는 데 의의가 있다.

Ⅰ. 서 론

비디오 기반 인간 행동 인식(HAR, Human Action Recognition)은 지능형 감시, 인간-컴퓨터 상호작용, 운전자 모니터링 등 다양한 응용에서 핵심적인 역할을 수행한다. 특히 I3D[1], SlowFast[2]와 같은 딥러닝 기반 모델과 HMDB51[3], UCF101[4] 등 대규모 벤치마크 데이터셋의 발전은 행동 인식 연구를 크게 진전시켰다. 그러나, 실제 응용에서는 특정 동작만 필요한 경우가 많으며, 불필요한 클래스와 배경 정보는 오히려 성능 저하 및 학습 혼선을 유발할 수 있다.

이에 본 연구에서는 HMDB51 벤치마크 데이터 셋의 51개 클래스 중 고정된 움직임을 포함하며 운전자 및 작업자 모니터링 응용에 적합한 11개 클래스를 선별하였다. 또한, Detectron2[5] 기반 전경 분리(마스크)를 적용하여 배경 제거가 행동 인식 성능에 미치는 영향을 검증하였다. 실험 결과, 전경 분리 데이터셋은 원본 대비 정확도가 전반적으로 낮았으며, 이는 전경 분리가 보편적 성능 향상 전략이 아님을 시사하며, 클래스 특성에 따른 선택적 활용이 필요함을 보여준다.

Ⅱ. 본론

2.1 클래스 서브셋

본 연구에서는 HMDB51의 51개 행동 클래스 중 실제 생활 맥락과 연관성이 높은 eat, drink, smile, laugh, talk, smoke, clap, wave, shake_hands, pick, pour의 11개 클래스를 선별하여 서브셋을 구축하였다. 각 클래스는 훈련 세트 70개, 테스트 세트 30개로 구성되며, 전체 데이터셋은 훈련 770개, 테스트 330개로 재편성하였다.

2.2 전경 분리 및 행동 인식 아키텍처

사람 중심의 동작 인식을 위해 Detectron2의 Mask R-CNN[6]을 적용하였다. 각 프레임에서 사람 객체를 검출하여 전경 분리 영상을 생성하고, 배경 영역을 제거함으로써 전경 중심의 데이터셋을 구축하였다. 결과적으로, 동일한 비디오에 대해 원본 데이터셋과 전경 분리 데이터셋의 두 버전을 생성하였다. 또한, 행동 인식 모델로 SlowFast 네트워크를 사용하였다. SlowFast는 저주파 공간 특징을 학습하는 Slow pathway와 고주파 시간 특징을 학습하는 Fast pathway로 구성된 3D CNN 구조이다. 본 연구에서는 PyTorch Video[7]에서 제공하는 사전 학습된 모델을 사용하였다. RGB 비디오 클립을 16×8 프레임을 입력으로 사용하였으며, 모든 실험에서 동일한 하이퍼파라미터를 적용하여 전경 분리에 따른 성능 차이를 공정하게 비교하였다.

Ⅲ. 실험 및 결과

3.1 학습 설정

모델 학습은 SGD 옵티마이저와 dropout=0.5를 적용하여 30 epoch 동안 수행하였다. 배치 크기는 2로 고정하였다. 평가에는 test set을 사용하였으며, 지표는 Top-1 및 Top-5 정확도와 클래스별 정확도로 설정하였다. 실험은 원본 영상과 전경 분리 영상 간의 IoU를 산출하여 전경 분리 효과를 정량적으로 검증하고, 두 데이터셋을 각각 학습 및 평가하여 전경 분리가 행동 인식 성능에 미치는 영향을 분석하였다. 마지막으로, 클래스별 정확도를 비교하여 클래스 특성에 따른 성능 변화를 고찰하였다.

모든 실험은 Ubuntu 20.04 환경에서 Python 3.7, CUDA 11.1, PyTorch 1.8.0을, 하드웨어는 단일 NVIDIA TITAN RTX GPU를 사용하였다.

3.2 실험 결과

본 연구는 원본과 전경 분리 영상의 영역 일치도를 IoU 지표로 산출하여 배경 제거가 동작 정보 보존에 미치는 영향을 검증하였다. IoU는 모델 성능과 직접 연관되지는 않지만, 전경 분리의 적절성을 평가하는 지표로 활용될 수 있다. 그림 1은 11개 클래스의 원본(녹색)과 전경 분리(빨간색) 각각의 바운딩 박스에 대한 IoU 예시를, 그림 2는 클래스별 평균 IoU 결과를 제시한다. Smile, smoke, talk, laugh는 0.8 이상으로 안정적이었으나, drink, eat, pour는 소도구 소실로 인해 낮은 값을 기록하였다. Shake_hands는 다중 인물 검출 문제로 가장 낮은 결과를 보였다. 이는 전경 분리가 일부 클래스에서는 효과적이나, 특정 상황에서는 오히려 핵심 정보 손실을 초래할 수 있음을 시사한다.

그림 1. 클래스 별 전경 분리 영상과 원본 영상의 IoU

그림 2. 클래스별 평균 IoU

표 1은 원본과 전경 분리 데이터셋의 성능을 비교한 결과로, 원본은 Top-1 정확도 실험 2에서 88.18%, 전경 분리 데이터 셋은 실험 5에서 80.91%로 성능이 낮았다. 이는 배경 제거로 인해 행동 구분에 필요한 맥락적 단서가 손실된 결과로 해석되며, 단순히 전경만을 활용할 경우 행동 인식에 필요한 맥락적 단서가 제거되어 동작 구분이 어려워질 수 있음을 의미한다.

표 1. Top-1 정확도(%)

No.	Dataset	Class	LR	WD	Top-1	Top-5
1	Original	51	1e-3	5e-4	77.10	95.28
2	Original	11	1e-3	5e-4	88.18	99.09
3	Original	11	1e-4	5e-4	75.15	97.27
4	Original	11	1e-3	1e-5	86.97	98.48
5	Mask	11	1e-3	5e-4	80.91	96.67

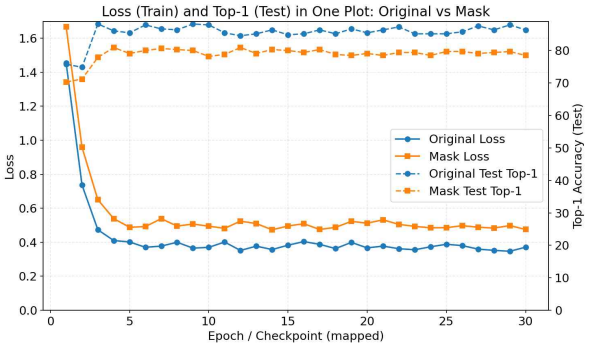

그림 3. Original과 Mask 영상의 학습 손실과 체크포인트 테스트 결과

그림 3과 그림 4는 각각 실험 2와 5에서의 학습 손실 및 checkpoint별 Top-1 정확도와 원본과 전경 분리된 영상에 대한 클래스별 정확도 변화를 나타낸다. 원본 데이터셋은 전경 분리보다 안정적으로 높은 성능을 보였으며, pick, shake_hands, smoke는 성능 차이가 없었으나, 다른 클래스들은 배경 단서 손실과 마스킹 오류 등으로 인해 성능이 저하되었다.

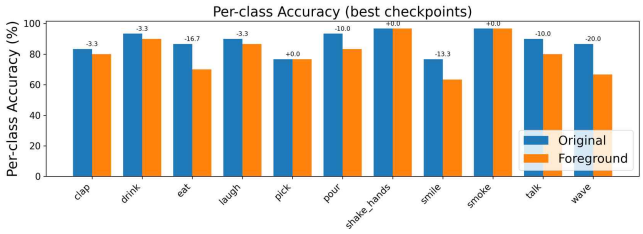

그림 4. 원본과 전경 분리 영상의 클래스별 Top-1 정확도

IV. 결론

연구는 HMDB51 데이터셋에서 선별한 11개 클래스를 대상으로, 원본 영상과 전경 분리 영상을 SlowFast 모델로 학습 및 평가하였다. HMDB51은 talk, laugh, smile 등 미묘한 표정 및 제스처 중심의 클래스가 다수 포함되어 있어, 본질적으로 클래스 간 구분이 어려운 데이터셋이다. 실험 결과, 전경 분리 데이터셋은 배경 정보 손실로 인해 원본 대비 낮은 성능을 보였다. 이는 행동 인식에서 전경과 배경이 상호보완적으로 작용함을 보여주며, 특히 HMDB51과 같이 시각적 유사성이 높은 클래스들에서는 단순 전경 분리만으로는 충분한 구분력을 확보하기 어렵다는 점을 시사한다. 따라서 향후 연구에서는 시간적 마스크 스무딩, 멀티모달 융합 등 전경과 배경 단서를 효과적으로 결합하는 전략이 요구된다.

ACKNOWLEDGMENT

This study was supported by the DGIST R&D Program of the Ministry of Science and ICT of Korea (25-IT-03).

참 고 문 헌

[1] Carreira, J.; Zisserman, “A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21 - 26 July 2017; pp. 6299 - 6308.

[2] Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. “SlowFast Networks for Video Recognition,” arXiv 2018, arXiv:1812.03982.

[3] Kuehne, H.; Jhuang, H.; Stiefelhagen, R.; Serre, T. “HMDB51: A Large Video Database for Human Motion Recognition,” In High Performance Computing in Science and Engineering ‘12; Nagel, W., Kröner, D., Resch, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 571 - 582.

[4] Soomro, K.; Zamir, A.R.; Shah, M. “UCF101: A Dataset of 101 Human Action Classes from Videos in the Wild,” CRCV-TR-12-01, November 2012.

[5] Yuxin Wu and Alexander Kirillov and Francisco Massa and Wan-Yen Lo and Ross Girshick, “Detectron2”, 2019, https://github.com/facebookresearch/detectron2.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. “Mask R-CNN”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969

[7] Fan, H.; Murrell, T.; Wang, H.; Alwala, K.V.; Li, Y.; Li, Y.; Xiong, B.; Ravi, N.; Li, M.; Yang, H.; et al. “PyTorchVideo: A deep learning library for video understanding,” In Proceedings of the 29th ACM International Conference on Multimedia (MM ’21), ACM: New York, NY, USA, 2021; pp. 3800 - 3803.