360 도 ERP 영상에서 행동 인식 모델 성능 향상을 위한 전처리 기법

박은수 [1], 유재성 [2], 김승환 [2], 류은석 [1*]
성균관대학교 [1], 가천대학교 [2]
espark804@skku.edu[1], poopoo96@gc.gachon.ac.kr[2],
whitekomani@gc.gachon.ac.kr[2], esryu@skku.edu[1*]

Preprocessing Methods for Action Recognition Model in 360-degree ERP Video

Eun-Soo Park[1], Jaesung Ryu[2], Seunghwan Kim[2], Eun-Seok Ryu[1*]
Sungkyunkwan University[1], Gachon University[2]

요 약

본 논문에서 Equirectangular projection(ERP) 영상을 행동 인식 모델에 입력하기전 제안하는 전처리를 통하여 성능을 향상시키는 것을 보인다. ERP 영상의 특성상 행동 인식을 하는데 불필요한 영역이 일반적인 2D 카메라로 촬영한 영상보다 많다. 또한 행동 인식은 사람이 Object of Interest(OOI)이다. 따라서 객체 인식 모델로 인간 객체를 인식한 후 Region of Interest(ROI)를 추출하여 불필요한 영역을 없애고, 왜곡 또한 줄어든다. 본 논문에서 제안하는 기법으로 전처리 후 CNN-LSTM 모델로 성능을 테스트했다. 제안하는 방법으로 전처리를 한 데이터와 하지 않은 데이터로 행동 인식을 한 정확도로 비교하였으며 제안하는 기법으로 전처리 한 데이터로 행동 인식을 한 경우 데이터의 특성에 따라 다르지만, 최대 61%까지 성능향상을 보였다.

1. 서론

최근 고성능 GPU 의 사용으로 기존 CPU 의 사용보다 처리 가능한 연산량이 대폭 증가함에 따라, 많은 연산이 필요한 딥러닝 관련 기술들이 연구가 활발히 진행되어 오고 있다. 딥러닝의 신경망 중 이미지 처리에 특화된 합성곱신경망(Convolution Neural Network, CNN)을 이용한 여러 기법의 발달과 함께, 객체 인식, 행동 인식, 이미지 캡셔닝 등과 같은 딥러닝이 적용된 이미지 프로세싱 연구가 빠른 속도로 진행되어 오고 있다[1]. 이 중에서 행동 인식 관련 연구는 여러 방향을 통하여 활발히 진행되어 오고 있으며, 객체 인식에 비하여 어려운 주제로 알려져 있다.

최근 가상현실(Virtual reality, VR) 관련된 데이터 즉, Salient 360[2], Sports-360[3]과 같은 360 영상 데이터가 많이 배포되고 있다. 그에 따라 머리에 장착 가능한 영상 제공 장치인 head-mounted display(HMD)와 360 영상 데이터를 취득할 수 있는 360 카메라가 시장에 보급되고 있다. 이러한 여러 장비에서 사용자가 이질감을 느끼지 않을 정도의 재생 속도를 제공하려면 낮은 지연 속도와 Ultra-high-definition(UHD) 이상의 초 고화질 360 영상이 요구된다[4]. 이와 같은 요구사항을 해결하기 위하여 수많은 연구가 진행되어오고 있는데, 서론에서 몇 가지를 소개하도록 한다. (i) 비대칭 코어 프로세싱 기반 타일 분할 및 할당 시스템[5-7]. (ii) 타일 기반 Motion-constrained tile set(MCTS)[8-10]. (iii) 기존보다 더 적은 수의 디코더와 더 적은 대역폭을 요구하는 방법[11]. (iv) 카메라의 위치에 따른 우선순위를 적용하여 비균등 다운 샘플링을 적용한 대역폭 절감 방법[12, 13]. (v) 영상의 프로젝션 포맷 변경을 통한 대역폭 절감 방법 이때 프로젝션 포맷은 360 영상을 2D 상에서 나타내기 위한 기법으로 ERP, Cube mapping(CMP) 등이 있다[14].

본 논문에서는 ERP 영상 데이터를 행동 인식 모델이 입력하기 전에 본 논문에서 그림 1 과 같이 제안하는 전처리 방법을 적용하여 산출된 행동 인식 성능과 ERP 영상 데이터에 전처리를 하지 않은 행동 인식 성능을 비교하여 성능 향상이 이루어 졌는지 실험 및 분석한다.

그림 1 제안하는 관심 영역 추출을 이용한 행동 인식 전처리 기법

본 논문은 2 장에서 행동 인식과 관련된 여러 관련 연구들을 소개한다. 3 장에서 제안하는 ERP 영상에서 행동 인식 모델 성능 향상을 위한 전처리 기법에 관한 것을 서술한다. 4 장에서 전처리 후 테스트한 행동 인식과 전처리 하지 않은 데이터를 입력한 동일한 모델을 비교하여, 행동 인식 성능에 관한 실험 및 분석을 한다. 5 장에서 결론 및 향후 연구를 서술한다.

2. 관련 연구

행동 인식을 할 수 있는 데이터는 행동의 특성상 시간적 성분을 갖고 있는 것이다. 즉, 영상 데이터가 필요하다. 영상 데이터의 종류도 상당히 많은데, 그 예로 RGB-D, 2D 카메라 영상, 360 카메라 영상 등이 있다. 본 논문에서는 ERP 데이터를 이용한다. ERP 는 360 영상(구형 좌표)를 평면으로 변환한 것으로, 세계지도를 예로 들 수 있다. ERP 의 특징으로는 극점으로 갈수록 왜곡이 심해진다는 점인데, 이를 해결하고, CNN 모델에 ERP 데이터를 입력하여 객체인식을 하는 많은 연구들이 있었다. 기존 연구들은 다음과 같이 세가지 연구가 있다. (i) 360 영상을 ERP 로 변환하여 CNN. 모델에 입력하는 연구 (ii) 360 영상에서 타일단위로 추출하여 탄젠트 연산을 적용하여 왜곡을 완화시킨 후 모델에 입력하는 연구 (iii) 360 영상 특성을 이용하여 극점에서 수평 부분의 각도를 기준으로 커널 사이즈를 변형하여 앞서 언급한 2 가지 방법의 장점만을 얻으며 객체 인식의 성능을 높이는 연구가 있다[15].

3. ERP 영상에서 행동 인식 모델 성능 향상을 위한 전처리 기법

3 장에서 본 논문에서 제안하는 ERP 영상 입력 행동 인식 모델 성능 향상 기법에 관한 것을 서술한다. 1 절에서 제안하는 기법의 개요에 관한 설명을 하고, 2 절에서 ROI 추출기법에 관한 설명을 한다.

3.1 제안하는 전처리 기법 개요

본 논문의 연구 이전에도 많은 ROI 기반 행동 인식 연구들이 많았다[16-19]. 그러나 대부분의 연구들은 객체 1 개를 기준 또는 두 사람 간의 상호 작용을 연구를 한 것들이 많고, 데이터 셋 또한 사람이 인위적으로 만든 데이터 셋들을 사용하였다. 본 논문에서는 일반적으로 업로드되는 Youtube 에 있는 360 영상을 테스트한다.

본 논문에서 제안하는 전처리 기법의 기본 가정은 행동 인식은 인간이 OOI 이며, 인간의 주변 환경이 ROI 라는 가정을 전제로 한다.

딥러닝 신경망의 입력 사이즈는 모델에 따라 다르지만, 일반적으로 원본 이미지보다 작기 때문에 데이터를 입력할 때 다운사이징을 적용한다. 일반적인 ERP 영상은 기존 2D 카메라로 촬영한 영상보다 사이즈가 크거나 같기 때문에 원본 프레임을 다운사이징을 하여 모델에 입력할 경우 프레임의 특징, 즉 인간의 행동을 잘 표현하기 어렵다. 따라서 프레임 내의 ROI 를 먼저 추출한 후 행동 인식 모델에 입력하는 것으로 영상 내 특징을 기존 전처리를 진행하지 않았을 때 보다 더 효율적으로 추출할 수 있게 한다.

3.2 ROI 추출 기법

제안하는 전처리 기법을 적용한 행동 인식 과정은 그림 2 와 같다. 그림 3 은 UOF-101 데이터 셋의 클래스 중 하나인 Biking 을 Inference 하는 과정을 나타낸 것이다. 본 논문에서 객체를 인식하기 위하여 You only look once(YOLO)V3 모델을 이용하였다. YOLO 는 속도가 빠르지만, 다른 객체 인식 모델들에 비해 정확도가 뒤쳐지지 않아 선택하였다.

프레임 내의 person 객체만을 인식하기 위해 YOLO 소스를 person 객체만 인식하도록 수정하였다. 또한 객체 인식의 오답 가능성 여부가 있기 때문에 person 객체 인식 정확도가 90%이상일 경우에만 객체를 인식한 것으로 수정하였다.

입력된 프레임 내의 person 객체가 1 개일 경우 모델에서 인식한 좌표값을 이용하여 추출한다. 다중 객체일 경우 그림 2 와 같이 모든 인식된 person 객체의 상, 우 좌표는 최대값, 하, 좌 좌표는 최소값을 적용하여 최대 ROI 범위를 지정한다.

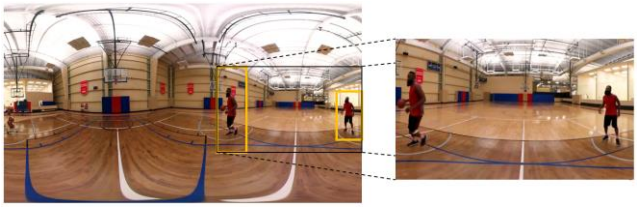

그림 2 프레임 내에 다중 객체가 존재할 경우 ROI 선정 방법

3.3 행동 인식 모델

본 논문에서 사용한 행동 인식 모델은 CNN-LSTM 모델이다. CNN-LSTM 모델의 장점은 CNN 의 연구가 진행되어 기존의 모델보다 더 성능이 좋은 모델이 나온다면 얼마든지 적용이 가능하다는 점이다. 단점으로는 inference 과정이 다소 어렵고 시간이 많이 소요된다. 본 논문에서 CNN 은 Inception V3 모델을 사용하였다. 그림 3 의 Inference 과정은 CNN 의 softmax 이전 단계인 average pool 레이어에서 특징을 추출한다. 추출된 feature 는 한 장의 프레임의 특징이므로 시간적 정보가 필요한 행동 인식의 경우 연속된 프레임들의 특징을 병합한다. 병합된 특징들은 시간적 정보를 포함하고 있는 데이터를 입력가능한 LSTM 에 입력한다. LSTM 의 산출물은 UCF-101 의 101 가지 클래스 중 1 가지가 출력된다.

3.4 데이터 셋

본 논문에서 이용한 데이터 셋은 UCF-101 데이터 셋으로 University of Central Florida 에서 제작한 데이터 셋이다. 총 13320 개의 비디오가 포함되어 있는 101 가지 인간의 행동에 관련된 영상들을 Youtube 에서 다운로드 받고 클래스 별로 구분하였다.

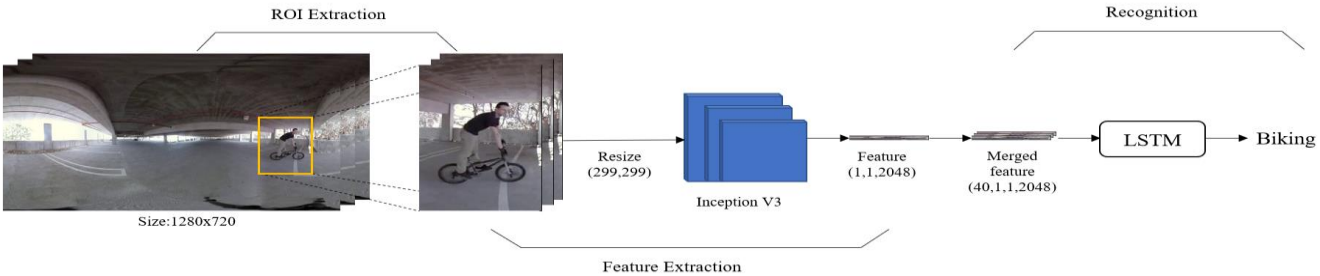

그림 3 제안하는 전처리기법을 적용한 행동 인식 모델 (CNN+LSTM) inference 과정

UCF-101 데이터 셋의 특징은 Youtube 에서 다운로드 받은 것이기 때문에 잡음이나 화면의 흔들림, 조도 등의 영향이 많다. 따라서 배우들이 연기한 데이터 셋들보다 현실적인 데이터 셋이라고 할 수 있다.

4. 행동 인식 실험 및 결과 분석

3 장에서 제안한 전처리 기법을 적용한 행동 인식 모델의 성능 실험을 한다. 실험 환경은 운영체제 Ubuntu 18.04 LTS, 그래픽 카드 GTX 1080TI 를 사용하였다. 언어는 객체인식 모델인 YOLO 를 수정할 때 C 언어를 사용하였고, 그 외 작업은 Python 2 버전과 3 버전을 사용하였다. 1 절에서 행동 인식 모델로 실험 및 결과를 보인다.

4.1 제안하는 전처리 기법을 적용한 행동 인식 실험 및 결과

테스트 데이터는 Youtube 에서 360 영상을 직접 다운로드 받았다. 테스트 영상의 기준은 UCF-101 데이터 셋의 클래스에 존재하는 행위를 하는지의 여부로 판단하였다. 테스트 데이터의 특징은 표 1 과 같다.

표 1 테스트 데이터의 특징

Sequence	Object	Camera movement	Number of frames
Biking 1	Multiple	O	1,000
Biking 2	Single	X	1,000
Basketball 1	Multiple	X	5,558
Basketball 2	Multiple	X	3,000
Bench press 1	Single	X	3,000

테스트 데이터들을 제안한 기법으로 ROI 를 추출한 후 UCF-101 데이터 셋으로 선행 학습된 CNN-LSTM 모델에 입력한다. 그림 4 는 ROI 를 추출한 실험 결과이다. 위에서부터 Biking 1, Biking 2, Basketball 1 이다.

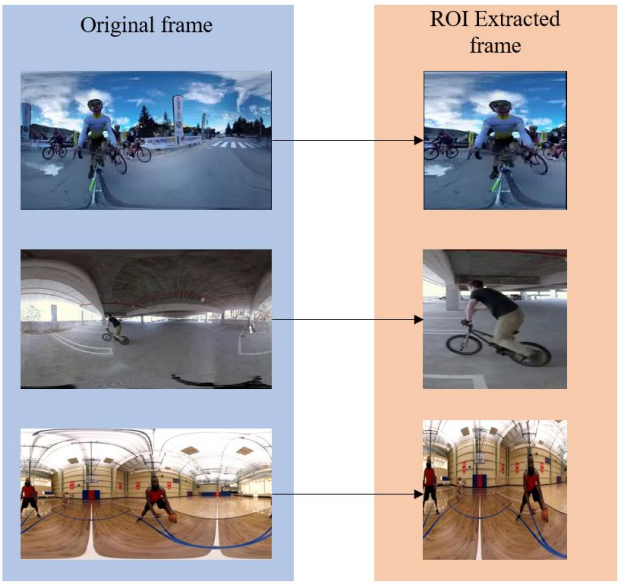

그림 4 테스트 데이터 셋에서 ROI 를 추출한 결과

CNN-LSTM 모델은 sequence 길이가 정의되어 있어야 한다. 다시 말해, CNN 으로 추출한 특징을 한 번에 얼마나 쌓을지를 정의해야 한다. 본 논문에선 40 으로 정의하였다. 평가 기준은 전체 프레임을 행동 인식 모델에 입력하고, 전처리를 하지 않고 inference 한 결과와 전처리를 하고 inference 한 결과를 비교한다. 이 때 정확도는 전체 프레임으로 평균을 낸 정확도이다. 실험 결과는 표 2 와 같으며 행동 인식 정확도가 최대 61%의 성능 향상을 보인다.

표 2 제안한 방법으로 전처리 후의 성능 증가 실험 결과

Sequence	Original	ROI Extracted
Biking 1	46%	62%
Biking 2	19%	80%
Basketball 1	12%	15%
Basketball 2	7%	20%
Bench press 1	14%	25%

4.2 결과 분석

표 2 와 그림 4 를 볼 때 Biking 2 의 경우가 가장 행동 인식 정확도 향상 폭이 큰데, 그 이유는 원본 데이터에서 행동 인식과 관련 없는 불필요한 부분이 상당히 많았다. 마찬가지로 Biking 1 이나, Basketball 1 과 같은 경우도 ERP 특성상 불필요한 부분과 왜곡된 부분이 많은데 ROI 추출로 인하여 불필요한 부분 그리고 왜곡된 부분이 제거되어 성능 향상이 이루어 진 것으로 보인다. Basketball 2 와 Bench press 1 의 원본 정확도가 낮은 이유는 Youtube 데이터 셋을 찾다 보니, 사람이 출현하지 않는 의미 없는 프레임들이 많기 때문이다.

5. 결론 및 향후 연구

본 논문은 360 영상을 ERP 로 변형하고, ERP 의 단점인 왜곡과 광범위한 불필요한 영역을 제거하기 위하여 ROI 추출 전처리를 적용한 후 행동 인식을 하는 것으로 성능 향상을 보였다. 데이터의 특성에 따라 다르지만 크게 61%의 성능 향상율을 보였으며 실험한 데이터 모두 성능이 낮아짐 없이 향상을 하였다. 향후 연구로 다중 객체가 프레임내에 존재할 경우 해당 행동을 하는 사람이 있고 관중이 있을 경우에 영상 내의 주체를 파악하는 등의 연구가 필요하다.

Acknowledgement

본 연구는 한국전력공사의 2016 년 선정 기초연구개발과제 연구비에 의해 지원되었음 (과제번호: R17XA05-68), 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2019-2017-0-01630)

참 고 문 헌

[1] Eun-Soo P., Seunghwan K., Jaesung R., Seondae K. Ghulam M., Eun-Seok R. "Action Recognition Reference Image Captioning," The Korean Institute of Broadcast and Media Engineers (KIBME) Summer Conference, pp.21-24, Jun. 19-21, 2019.

[2] J. Gutierrez, E. David, A. Coutrot, M. Perreira Da Silva, P. Le Callet, "Introducing UN Salient360! Benchmark: A platform for evaluating visual attention models for 360 contents," International Conference on Quality of Multimedia Experience (QoMEX),

Sardinia, Italy, May. 2018.

[3] Hou-Ning H., Yen-Chen L., Ming-Yu L., Hsien-Tzu C., Yung-Ju C., Min Sun. Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1396-1405. 2017.

[4] Mary-Luc C., Thomas S., Thierry F., Emmanuel T., Rob K. Quality Requirements for VR. 116th MPEG meeting of ISO/IEC JTC1/SC29/ WG11, MPEG 116/m39532. 2016.

[5] Hyun-Joon R, SungWon H, Eun-Seok R. "Prediction complexity based HEVC parallel processing for asymmetric multicores." Multimedia Tools and Applications 76, 23, pp.25271-25284. 2017.

[6] Hyun-Joon R, Bok-Gi L, Eun-Seok R. "Tile Partitioning and Allocation for HEVC Parallel Decoding on Asymmetric Multicores." The Journal of Korean Institute of Communications and Information Sciences (J-KICS), Vol.43, No.05, pp. 791-800. 2018.

[7] Seehwan Y, Eun-Seok R. "Parallel HEVC decoding with asymmetric mobile multicores." Multimedia Tools and Applications 76, 16, pp.17337-17352. 2017.

[8] Robert S, Yago S, Karsten S, Thomas S, Eun-Seok R, Jangwoo S. "Temporal MCTS Coding Constraints Implementation." 122th MPEG meeting of ISO/IEC JTC1/SC29/ WG11, MPEG 122/m42423. 2018.

[9] Jang-Woo S, Dongmin J, Eun-Seok R. "Implementing Motion-Constrained Tile and Viewport Extraction for VR Streaming." ACM Network and Operating System Support for Digital Audio and Video 2018 (NOSSDAV2018). 2018.

[10] Jang-Woo S, Eun-Seok R. "Tile-Based 360-Degree Video Streaming for Mobile Virtual Reality in Cyber Physical System." Elsevier, Computers and Electrical Engineering. 2018.

[11] Jong-Beom J., Soonbin L., Dongmin J, Il-Woong R., Tuan T. L., Jaesung R., Eun-Seok R." Implementing Multi-view 360 Video Compression System for Immersive Media ", The Korean Institute of Broadcast and Media Engineers (KIBME) Summer Conference, pp.139-142, Jun. pp.19-21, 2019.

[12] JongBeom J, Dongmin J, Jangwoo S, Eun-Seok R, "3DoF+ 360 Video Location based Asymmetric Down-sampling for View Synthesis to Immersive VR Video Streaming", MDPI, Sensors, 18(9):3148, Sep. 2018.

[13] JongBeom J., Dongmin J., Jangwoo S., Eun-Seok R., "Bitrate Efficient 3DoF+ 360 Video View Synthesis for Immersive VR Video Streaming", International Conference on ICT Convergence 2018 (ICTC2018), Sep. pp.17-19, 2018.

[14] JongBeom J., Dongmin J., Eun-Seok R., "3DoF+ 360 Video Projection Conversion for Saving Transmission Bitrates ", The Korean Institute of Broadcast and Media Engineers (KIBME) Fall Conference, Nov. pp.02-03, 2018.

[15] Su, Y. C., & Grauman, K. (2017). Learning spherical convolution for fast features from 360 imagery. In Advances in Neural Information Processing Systems (pp. 529-539).

[16] Jhuang, H., Gall, J., Zuffi, S., Schmid, C., & Black, M. J. (2013). Towards understanding action recognition. In Proceedings of the IEEE international conference on computer vision (pp. 3192-3199).

[17] Bregonzio, M., Li, J., Gong, S., & Xiang, T. (2010, September). Discriminative Topics Modelling for Action Feature Selection and Recognition. In BMVC (pp. 1-11).

[18] Berlin, S. J., & John, M. (2016, October). Human interaction recognition through deep learning network. In 2016 IEEE International Carnahan Conference on Security Technology (ICCST) (pp. 1-4). IEEE.

[19] Sharma, S., Kiros, R., & Salakhutdinov, R. (2015). Action recognition using visual attention. arXiv preprint arXiv:1511.04119.