360도 비디오에서 관심 영역 추출을 통한 행동 인식 성능 향상 기법

저자
(Authors)

박은수, 유재성, 류은석

출처
(Source)

한국통신학회지(정보와통신) 36(9), 2019.8, 8-16(9 pages)

The Journal of The Korean Institute of Communication Sciences 36(9), 2019.8, 8-16(9 pages)

발행처
(Publisher)

한국통신학회
Korea Institute Of Communication Sciences

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09221763

APA Style

박은수, 유재성, 류은석 (2019). 360도 비디오에서 관심 영역 추출을 통한 행동 인식 성능 향상 기법. 한국통신학회지(정보와통신), 36(9), 8-16

편집위원 : 김중헌(고려대학교)

360도 비디오에서 관심 영역 추출을 통한 행동 인식 성능 향상 기법

박은수, *유재성, 류은석
성균관대학교, *가천대학교

요 약

인공지능과 360 영상 관련된 연구들이 발전하면서 360 영상을 입력으로 하는 행동 인식 관련 연구가 많이 이루어지고 있다. 본 논문에서는 360 영상을 등장방형도형 (ERP)으로 변환 후 행동 인식 모델에 입력하였을 경우 2D 영상에 비해 인식률이 낮아지는 현상을 해결하기 위한 처리 과정을 제안한다. (i) 객체 인식을 통하여 행동 인식의 주체가 되는 인간 객체를 인식. (ii) 인식된 객체의 좌표들을 이용하여 관심 영역 (ROI) 생성. (iii) 생성된 관심 영역을 크롭 (Crop). 위 과정을 거쳐 생성된 이미지를 학습된 행동 인식 모델에 입력하여 기존의 전처리 과정이 없는 360 영상 입력 행동 인식 정확도 보다 최대 61%가 높아짐을 보인다. 또한 객체 인식된 프레임만을 추출하여 행동 인식할 경우 영상 요약의 효과도 볼 수 있으며 최대 68%의 더 높은 행동 인식 정확도를 얻을 수 있다.

Ⅰ. 서 론

최근 고성능 GPU의 사용으로 처리 가능한 연산량이 대폭 증가함에 따라, 많은 연산이 필요한 딥 러닝 관련 기술들이 연구가 활발히 진행되어 오고 있다. 이미지 처리에 특화된 합성곱신경망(Convolution Neural Network, CNN)을 이용한 여러 기법의 발달과 함께, 객체 인식, 행동 인식, 이미지 캡셔닝 등과 같은 이미지 프로세싱 연구가 상당히 빠른 속도로 진행되어 오고 있다[1]. 이 중에서 행동 인식 관련 연구는 여러 방향을 통하여 활발히 진행되어 오고 있으며, 상당히 어려운 주제로 알려져 있다.

행동 인식 관련된 연구는 2가지의 목적이 있다. (i) 인간의 행위 의도를 인식하고, 그에 맞는 서비스를 제공하는 컴퓨터 비전 또는 인공지능 측면에서의 목적 (ii) 컴퓨터나 로봇이 인간과 유사한 방법으로 의사를 표현할 수 있도록 하기 위한 Human-computer interface(HCI). 본 논문에선 인간의 행동을 인식 및 분류하는 (i) 과 목적이 동일하다[2].

행동 인식은 영상 데이터를 입력으로 하기 때문에 많은 양의 데이터를 갖는다. 또한, 영상 데이터는 영상이 재생되는 시간 동안 프레임들의 변화되는 정보를 기반으로 하는 학습이 중요하다. 영상 데이터는 2D 영상, 360 영상 등이 있다. 최근 가상현실(Virtual reality, VR) 관련된 데이터 즉, Salient 360[3], Sports-360[4] 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[5]. 이와 같은 요구사항을 해결하기 위하여 수많은 연구가 진행되어오고 있는데, 서론에서 몇 가지를 소개하도록 한다. (i) 비대칭 코어 프로세싱 기반 타일 분할 및 할당 시스템[6][7][8]. (ii) 타일 기반 Motion-constrained tile set(MCTS)[9][10][11]. (iii) 기존보다 더 적은 수의 디코더와 더 적은 대역폭을 요구하는 방법[12]. (iv) 카메라의 위치에 따른 우선순위를 적용하여 비균등 다운 샘플링을 적용한 대역폭 절감 방법[13][14]. (v) 영상의 프로젝션 포맷 변경을 통한 대역폭 절감 방법 이때 프로젝션 포맷은 360 영상을 2D 상에서 나타내기 위한 기법으로 등장방형도법 (Equirectangular projection, ERP), Cube mapping(CMP) 등이 있다[15].

이처럼 딥 러닝을 이용한 행동 인식 연구들이 많이 진행되고 있는 가운데, 360 영상 데이터도 많이 배포되고 있다. 위에 기술한 상황 때문에 360 영상 데이터를 입력으로 하는 행동 인식 관련 연구가 현재는 잘 찾아볼 수 없으나, 앞으로 많은 연구가 진행될 것으로 보인다.

본 논문에서는 〈그림 1〉과 같이 360 영상 데이터를 입력으로 하는 행동 인식을 진행한다. 360 영상 데이터는 2D 영상과 비교하여 비교적 넓은 범위의 영상을 가지기 때문에 딥 러닝 모델의 특성상 영상에서 특징을 잘 추출하는 것이 어렵다. 따라서 전처리 과정으로 객체 인식 모델을 사용하여 관심 영역(Region of Interest, ROI)을 추출하여 기존 전처리 없이 진행된 360 영상 데이터의 행동 인식보다 높은 성능을 도출한다.

본 논문은 2장에서 행동 인식을 하기 위한 여러 가지 연구를 딥 러닝 이용 여부에 따라 나누어 설명하였다. 3장에서 본 논문에서 제안하는 모델에 대해 자세히 기술한다. 4장에서 제안하는 모델로 여러 가지 실험을 한 내용을 기술한다. 마지막으로 5장에서 결론 및 향후 연구에 관해 기술한다.

Ⅱ. 관련 연구

2장에서 행동 인식과 관련된 수많은 연구를 딥 러닝 이용 여부에 나누어 서술한다. 1절에서 딥 러닝 이전의 행동 인식 기법에 관해 서술한다. 2절에서 딥 러닝을 사용한 여러 기법 및 모델들을 설명한다.

그림 1. 제안하는 360 영상 데이터를 입력으로 하는 행동 인식 성능 향상 기법
Fig. 1. Action recognition performance improvement technique with 360 video data

1. 딥 러닝을 이용하지 않은 행동 인식 기법

과거에는 행동 인식을 하기 위하여 실루엣을 사용하였다. 그러나 실루엣을 추출하였을 때의 결과물은 실루엣만을 완벽하게 추출하는 것은 다소 어렵기 때문에 잡음(Noise)이 심하다. 따라서 잡음을 줄이는 방법으로 범위 지정이나, 윤곽(Contour) 기법 등을 적용하는 연구가 많이 진행되었다. 그에 따른 연구로는 (i) 단일 이미지에서 실루엣을 추출하고, 후속 다중 이미지에서 프레임별 차이점들을 모아 행동 인식을 하였다. 결과물은 Motion energy image(MEI)와 Motion history image (MHI) 이며, MEI는 모션이 어디서 발생하였는지에 대한 정보를 표시해 주며, MHI는 실루엣의 모션 강도를 표시해준다[16]. (ii) 스타 스켈레톤(Star-skeleton)에 컨투어를 적용하여 신체 부위를 나누어 행동 인식을 하는 연구가 진행되었다. 이때 스타 스켈레톤은 폴리곤 표현 방식 중 하나이다[17]. (iii) 시간에 따른 실루엣의 차이 즉, 거리를 유클리디언 거리(Euclidean distance)계산법을 이용, 각 프레임의 차이를 계산하여 행동 인식을 하는 연구가 진행되었다[18]. 이외의 실루엣을 이용하지 않는 방법들은 (i) 사람 객체 추적 기술을 적용하여 추출된 관심 영역을 입력으로 하는 행동 인식을 한 연구로, 스포츠 데이터 셋을 이용하였다. 한 가지 흥미로운 점은 추출된 객체가 상당히 작은 점인데 성능이 좋게 나온 것이다[19]. (ii) 광학 흐름(Optical flow)에서 키네틱(Kinetic) 특징을 추출하였다. 키네틱 특징은 분기(divergence), 와도(Vorticity), 대칭(Symmetry) 등의 특징을 나타낸다. 이때 이 특징들을 주성분 분석(Principal Component Analysis, PCA)을 통해 키네틱 모드를 분석할 수 있다[20].

2. 딥 러닝을 이용한 행동 인식 기법

현재의 행동 인식 연구들은 고성능 GPU를 사용한 머신 러닝 기법을 사용하여 인간의 행동을 분석하는 방법이 기존의 룰 베이스(Rule-base) 기법들 보다 대체로 성능이 높게 나오기 때문에 머신 러닝을 사용한 연구가 많이 이루어지고 있다. 그에 따른 연구로는 (i) 객체의 시간에 따른 유사도를 나타내는 피셔 벡터 (Fisher vector)를 잘 추출한다는 이점을 가진 궤적 추출 기법 (Trajectory extraction) 그리고 자동으로 이미지의 특징을 추출해주는 깊은 신경망 중 하나인 합성곱신경망의 장점을 취하여 성능을 높이는 연구가 있다[21]. (ii) 입력 데이터 타입을 공간 정보를 담은 이미지 데이터, 시간 정보를 담은 광학 흐름 데이터로 나누어 두 방향의 합성곱신경망을 이용하여 그 결과를 퓨전 (fusion)하여 행동 인식을 하는 연구. 많은 행동 인식 관련 연구들이 참고 문헌[22]을 바탕으로 진행되었다[22]. (iii) 사람 객체를 스켈레톤 데이터로 변환하여 몸통(머리 포함), 팔, 다리 각 두 파트 씩 총 5파트로 나눈 후 각각의 파트를 입력 데이터로 갖는 개량된 순환신경망(Recurrent Neural Network, RNN)을 사용하여 행동 인식하는 연구[23]. (iv) 스켈레톤으로 이루어진 데이터 셋을 이용하여 각 시퀀스 별 거리 특징 (Features of distance position) 데이터는 Long Short-Term Memory (LSTM)에 입력하고, 공동 거리 지도(Joint distance map)는 데이터 형태 (Type)이 이미지이기 때문에 합성곱신경망 (CNN)에 입력하는 연구[24]. (v) 직접 센서 측정 방법으로 인간의 행동 별 자료를 수집, 분석하고, 인공지능과 접목하여 인공지능 기반의 방법론에 대한 문제점들을 보완한 연구[25]. (vi) 심층 신경망의 깊이에 따른 성능 분석[26]. (vii) 강화학습을 이용하여 순환신경망의 파라미터를 최적화한 후 스켈레톤 데이터를 이용하여 행동 인식을 하는 연구[27].

즉, 과거의 행동 인식 기법은 이미지에서 데이터를 추출하여 성능을 높이는 기법을 주로 사용하였고, 최근 들어서는 센서 및 여러 키넥트 등의 장비로 추가 정보를 이용하여 성능을 높이는 연구가 진행되고 있다.

위 소개된 관련 연구들은 대부분 2D 영상 및 추가 데이터 (e.g. 스켈레톤, 센서 데이터 등)를 이용한 연구들이 주를 이루고 있다. 현재 360 영상에 대한 행동 인식 관련 연구는 잘 찾아보기 힘들다.

Ⅲ. 객체 인식을 활용한 관심 영역 추출 기법 적용 행동 인식

본 논문에서는 360 영상 데이터를 입력으로 받아 행동 인식을 진행할 때 전처리 과정인 객체 인식 모델을 이용하여 영상의 관심 영역을 추출하여 성능을 높이는 방법을 제안한다.

1. 기존의 360 영상 관심 영역 추출 기법의 한계

360 영상에서 관심 영역을 추출하기 위한 연구로 Deep 360[4]이 있다. Deep 360은 영상에서 프레임을 추출하여 객체 인식을 한 후 이미 학습된 Selector RNN을 이용하여 관심 객체를 선택한다. 마지막으로 이미 학습된 Regressor RNN을 이용하여 Natural field of view (NFoV)를 생성한다. RNN은 수작업으로 프레임들에 객체의 좌표를 입력한 데이터 셋을 이용하여 학습하였다. 위 연구에서 한계점은 클래스 별로 가중치가 필요하다는 점이다. 클래스는 총 5가지 (농구, 파쿠르, 자전거, 스케이트보드, 춤)를 이용하였다. 앞서 소개한 연구에서의 문제점을 돌파하고자 본 논문에서 다음과 같은 360 영상 입력 행동 인식 처리 과정을 제안한다.

2. 360 영상 입력 행동 인식 처리 과정

제안하는 처리 과정은 〈그림 2〉와 같다. 360 동영상을 프레임 단위로 분해하여 각 프레임을 객체 인식 모델에 입력한다. 본 논문에서 사용한 객체 인식 모델은 You only look once (YOLO)이다. YOLO는 1 stage 객체 인식 모델로 속도가 빠르고, 성능은 다른 State of the arts (SOTA) 객체 인식 모델들과 비교하였을 때 정확도가 크게 뒤처지지 않는 것이 장점인 모델이다.

본 논문에서 객체가 인식된 이미지에서 관심 영역을 추출하기 위하여 알고리즘 1과 같은 방법을 사용하였다.

알고리즘 1
Algorithm 1
객체 인식을 통하여 관심 영역을 추출해내는 과정
The process of extracting the region of interest through object recognition

```
I: Frame
YOLO: Object detection model
Bi: Information of boxes
F: Flag
Nb: Detected number of boxes
L, R, T, B: Coordinates of the box
Oc: detected class
Op: detected accuracy

Bi, Nb ← YOLO(I)
F ← 0
for i < Nb do
if Oc = "person" and Op >= 0.9 then
L, R, T, B ← Bi
add_array(L.R.T,B)
F ← 1
end if
end for
if F = 1 then
  R, B ← max(R, B)
  L, T ← min(L, T)
  I ← crop_image(I, L, R, T, B)
  save_image(I)
else
  save_image(I)
end if
```

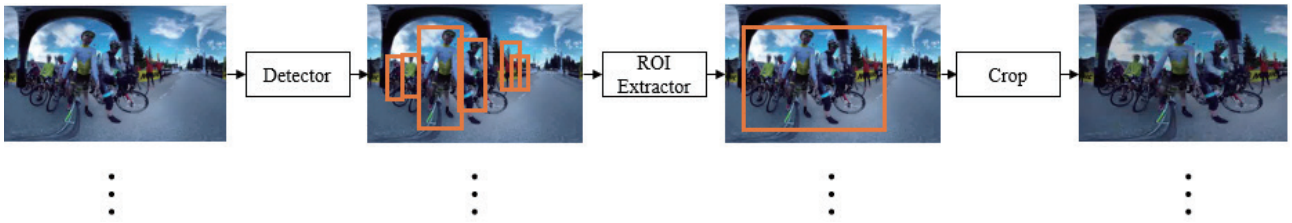


그림 2. 제안하는 360 영상 입력 행동 인식 처리 과정 (a) 원본 영상에서 추출된 단일 프레임 (b) 객체 인식 프로그램으로 "person" 객체만 인식된 이미지 (c) 인식된 바운더리 박스 좌표를 기반하여 관심 영역을 추출한 이미지 (d) 추출된 관심 영역의 좌표로 크롭한 이미지
Fig. 2. 360 video input action recognition process (a) Single frame extracted from original image (b) Image that only "person" object is recognized by object recognition program (c) Image extracted region of interest based on recognized boundary box coordinates (d) Cropped image with the coordinates of the extracted region of interest

먼저 객체 인식 모델인 YOLO에 360 영상에서 추출한 프레임을 입력하여 모든 클래스의 객체를 인식한다. 행동 인식에 필요한 관심 객체 (Object of interest, OOI)는 사람 객체이다. 따라서 객체 인식 모델은 전체 이미지에서 인식되는 클래스 (Class) 중 조건문을 넣어 행동의 주체가 되는 사람. 즉, "person" 객체만 인식하도록 수정하였다. 또한 오 인식률(False acceptance rate, FAR)을 줄이기 위하여 정확도가 90% 이상일 때에만 인식되도록 수정하였다. 한 프레임 안에 "person" 객체가 2개 이상일 수 있기 때문에 좌표들을 배열에 넣어 저장한다. 이후 이미지 내에 "person" 객체가 존재한다면, 영상의 (0, 0) 좌표가 좌상단에 있기 때문에 저장된 배열 내의 값 중 가장 큰 값을 저장하는 것은 오른쪽 좌표, 하단 좌표이다. 가장 작은 값을 저장하는 것은 좌측 좌표, 상단 좌표이다. 이후 산출된 좌표들을 이용하여 영상을 자른다 (Crop). "person"객체가 인식이 되지 않았다면 원본

프레임을 저장한다. 마지막으로 전 처리된 영상을 행동 인식 모델에 입력한다.

3. 행동 인식 모델

3절에서 본 논문에서 이용한 행동 인식 모델에 관해 기술한다. 1항에서 행동 인식 모델의 개요와 여러 행동 인식 모델의 종류에 관해 설명한다. 2항에서 학습 과정에 관해 설명한다.

3.1 행동 인식 모델 개요

행동 인식 모델은 합성곱신경망만을 이용한 것, CNN-RNN 모델, Convolution 3D 등이 있다. 이 중 실험 결과 성능이 가장 높았던 합성곱신경망 (CNN)과 LSTM을 병합하여 만든 모델을 본 논문에서 이용한다. CNN-LSTM 모델의 이점은 합성곱신경망을 유동적으로 변경할 수 있는데 Inception[28], ResNet[29] 등 발전해 나아가는 합성곱신경망에 맞추어 변경하면 성능 또한 높아진다. 그러나 단점으로는 학습 과정이 다른 모델들에 비해 복잡하다.

3.2 CNN-LSTM 모델 학습 과정

먼저 합성곱신경망은 2D영상 데이터를 이용하기 때문에 영상 전체를 입력할 수 없다. 따라서 영상을 프레임 단위로 분해해야 한다. 본 논문에서 OpenCV모듈을 이용하여 프레임을 추출하였다. 추출된 프레임을 합성곱신경망에 입력하기 전 합성곱신경망의 모델을 정의해야 한다. 본 논문에서 Inception V3 모델을 사용하였다. Inception V3모델은ImageNet 데이터베이스의 1백만 개가 넘는 이미지로 사전 훈련된 합성곱신경망이다. 합성곱신경망을 사전 훈련하지 않고 학습을 진행하려 하면 상당히 많은 시간이 소요되고 성능도 좋지 않기 때문에 사전 훈련된 신경망을 이용한다. 이미지를 입력하고 레이어들을 거쳐 입력한 이미지의 특징을 가장 잘 나타내는 합성곱신경망의 마지막 레이어에서 특징값 (Feature)을 추출한다. 위 과정에서 추출된 특징 데이터는 단일 이미지의 특징 데이터이다. 행동 인식은 시간적 정보가 필요한 기술로 여러 프레임을 한 번에 입력하여 처리한다. 따라서 순차적 프레임의 특징 데이터를 이미 정해진 길이만큼 쌓는다 (Append). 합쳐진 특징 데이터는 LSTM으로 입력되어 학습을 시작한다.

3.3 데이터 셋

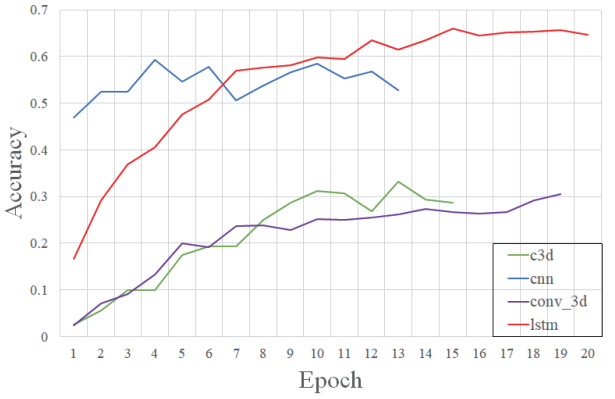
행동 인식에 사용된 데이터 셋은 행동 인식에 가장 많이 사용되는 UCF-101을 사용하였다. UCF-101 데이터 셋은 University of Central Florida에서 제작한 데이터 셋이다. YouTube에서 다운받은101 가지 인간의 행동에 관련된 데이터 셋이 포함되어 있다. 총 13320개의 비디오가 포함되어 있으며, 다양한 행동, 가변적인 카메라 움직임, 다양한 오브젝트 등을 포함하고 있다. 또한 여러 데이터 셋들은 현실과는 다른 준비된 영상들을 포함한 데이터 셋들이지만, UCF-101은 YouTube에서 다운로드 받아 직접 분류를 하였기 때문에 현실적인 요소를 포함하고 있다.

Ⅳ. 실험 및 결과 분석

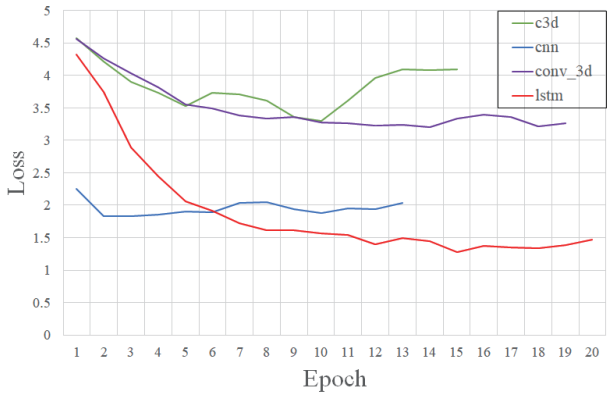
4장에서 본 논문에서 제안한 기술을 적용한 실험 및 부가 실험들을 진행한다. 실험 환경은 Ubuntu 18.04 LTS 운영체제, GTX 1080TI 및 RTX 2080TI를 이용하였다. 언어는 YOLO를 수정할 때 C언어를 이용하였고 그 외로는 Python 2버전과 3버전을 이용하였다. 1절에서 행동 인식 모델들을 UCF-101 데이터 셋으로 학습시킨 후 가장 성능이 좋은 모델을 검증하는 실험을 진행한다. 2절에서 추가적인 성능을 높이기 위해 전처리과정 중 하나인 히스토그램 평활화 기법 적용에 대한 실험을 진행한다. 3절에서 제안한 기술을 적용한 360 영상 관심 영역 추출 기반 행동 인식에 대한 실험을 진행하고 분석한다.

1. 행동 인식 모델 실험 및 결과

본 논문에서 이용한 행동 인식 모델은 CNN-LSTM모델이었다. 이는 다른 모델들 보다 높은 UCF-101 테스트 데이터에 대한 정확도를 가지기 때문이다. 이에 대한 실험을 진행하였다. 모델은 CNN-LSTM, C3D[30], conv3d, CNN을 실험하였으며 CNN을 제외한 모델의 프레임 길이는 40으로 맞추었다. 앞서 언급한 모델들 중 C3D와 conv3d 모델의 차이는 conv3d가 C3D 모델을 기반하여 하이퍼파라미터 및 모델을 변형했다. 실험 결과는 그림 3 및 표 1과 같다. CNN-LSTM의 정확도가 64%이고, 손실 값이 다른 모델들 보다 가장 낮아지는 것으로 보아 다른 행동 인식 모델들 보다 성능이 가장 높음을 알 수 있다.



(a)



(b)

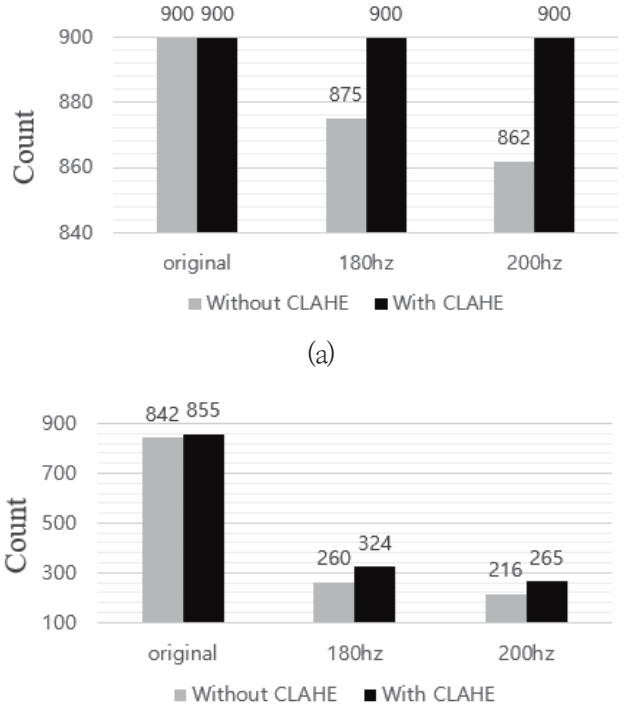
그림 3. 여러가지 행동 인식이 가능한 모델로 학습 후 테스트 시 측정한 정확도와 손실 값 그래프 (a) 모델들의 테스트 정확도를 측정한 그래프 (b) 모델들의 테스트 손실 값을 측정한 그래프
Fig. 3. Graph of accuracy and loss values measured after testing with various action recognition models (a) Graph of test accuracy of models (b) Graph of test loss values of models

표 1. UCF-101 데이터 셋을 이용한 행동 인식 모델 실험 결과
Table 1. Experimental results of action recognition model using UCF-101 dataset

Type of model	Accuracy	Loss
C3D	0.29	4.09
CNN	0.52	2.03
Conv3d	0.30	3.26
CNN-LSTM	0.64	1.47

2. 히스토그램 평활화 적용 실험 및 결과

본 논문에서 행동 인식 모델의 테스트를 위해 사용한 데이터는 YouTube에서 다운로드 받은 360 영상 데이터이다. 테스트 영상은 조도의 영향 및 화질의 고저가 있다. 앞서 말한 요인들을 해결하기 위해 적용할 수 있는 기술은 히스토그램 평활화 기법이 있다. 히스토그램 평활화 기법은 BBHE[31], RMSHE[32], DSHIHE[33], RSIHE[34], CLAHE[35] 등과 같이 여러 기법들이 있다. 이 중 가장 원본 훼손이 적은 Contrast limited adaptive histogram equalization (CLAHE) 기법을 적용한다. CLAHE 기법이 객체 인식 성능을 향상시키는지에 대한 여부를 그림 4와 같이 실험해보았다. 테스트한 데이터 셋은 Living Room Dataset[36]이며 컴퓨터 모니터와 책상 등 여러 오브젝트가 존재하는 영상이다. 또한 조도 및 카메라 움직임 등이 추가되어 있다. 테스트는 각 클래스 (원본, 180hz, 200hz)별로 900장씩 프레임을 추출하여 테스트하였다. 테스트하는 데이터 셋 영상에서 가장 큰 그리고 뚜렷한 객체는 모니터이다. 따라서 테스트시 모니터만을 인식하도록 설정하고 원본과 잡음 현상이 있는 영상을 테스트하였다. 그림 4에서 주파수 (Hz)는 높을수록 잡음 현상이 높다는 것을 의미한다. 실험 결과는 원본 영상 또는 잡음 현상이 있는 영상에서 F-RCNN보다 YOLO V3로 객체 인식한 성능이 더 좋았다. 또한 CLAHE를 적용한 영상에서 두 객체 인식 모델 모두 성능이 높아진 것을 확인할 수 있었다. 위와 같은 현상은 CLAHE의 적용 효과와 관련이 있다. CLAHE는 명암 대비(Contrast)를 제한하여 블록현상이 많이 발생하지 않도록 조절이 가능하여 원본의 훼손이 적기 때문에 인식률의 저하가 발생하지 않는 것이다. 또한 히스토그램 평활화를 하였기 때문에 명암 대비가 높아져서 객체의 윤곽선 등 영상이 선명해지는 효과가 있기 때문에 성능이 높아진 것으로 보인다.



(a)

(b)

그림 4. 원본 영상과 잡음현상이 있는 영상에 YOLO V3와 F-RCNN으로 객체 인식할 경우 CLAHE의 적용 효과 실험 결과
(a) YOLO V3로 객체 인식을 한 결과 (b) F-RCNN으로 객체 인식을 한 결과
Fig. 4. Experimental results of the application of CLAHE for object detection with YOLO V3 and F-RCNN on original video and video with noise (a) Result of object detection with YOLO V3 (b) Result of object recognition with F-RCNN

3. 360 영상 관심 영역 추출 실험 및 결과

3장 2절에서 기술한 방법으로 360 영상에 대한 관심 영역 추

출을 실험하였다. 테스트 데이터는 YouTube에서 다운로드 받은 360 영상으로 테스트를 진행하였다. 테스트 영상의 다운로드 받는 기준은 "person" 객체가 존재할 때와 객체가 움직임이 있을 때를 기준으로 선정하였다. 또한 학습한 데이터인 UCF-101 데이터 셋의 클래스에 속하는 행동을 하는 영상을 기준으로 하였다. 실험 결과는 그림 5와 같다. 그림 5를 보면, 360 영상 즉 ERP 프레임에서 행동 인식에 사용되지 않는 부가 영역들이 제거되고 인간을 위주로 관심 영역이 추출되는 것을 볼 수 있다.

그림 5. 원본 360 영상 프레임에서 관심 영역을 추출한 결과
(a) 원본 프레임 (b) 관심 영역을 추출한 프레임
Fig. 5. Results of extracting regions of interest from original 360 video frames
(a) Original frame (b) Frame extracted region of interest

4. 360 영상 관심 영역 추출 기반 행동 인식 실험 및 결과

본 논문에서 제안한 360 영상에 객체 인식으로 관심 영역을 추출 후 행동 인식을 하는 것에 대한 실험을 진행하였다. UCF-101로 학습할 때 사용한 영상 데이터의 길이가 100이하인 것도 있었기 때문에 프레임 길이를 지정할 때 10, 20, 30, 40, 50프레임 단위로 학습하였다. 따라서 테스트도 테스트 영상을 10에서부터 50프레임까지 지정한 후 분해하여 진행한다. 프레임 길이 설정도 행동 인식 정확도에 영향을 끼치기 때문에 제안하는 처리과정을 거친 360 영상 입력 행동 인식 모델의 결과도 그림 6과 같이 실험을 진행해 보았다. 큰 차이는 없었으나, 프레임의 길이 설정에 따라 정확도 및 손실 값이 달라지는 것을 확인할 수 있었다. 앞서 언급한 것과 같이 영상의 길이가 100프레임 이하인 경우도 있기 때문에 프레임 길이를 50으로 지정한다면 두 번째 50개의 프레임을 취합할 때 50 프레임 이하가 되므로 버려진다. 따라서 영상을 절반밖에 사용하지 못하는 상황이 발생한다. 그렇기 때문에 프레임 길이가 너무 길어도 안되며, 너무 짧아도 시간적 정보를 충분히 얻을 수 없기 때문에 높은 행동 인식 모델의 성능을 위해서 적당한 프레임 길이를 설정해야 한다.

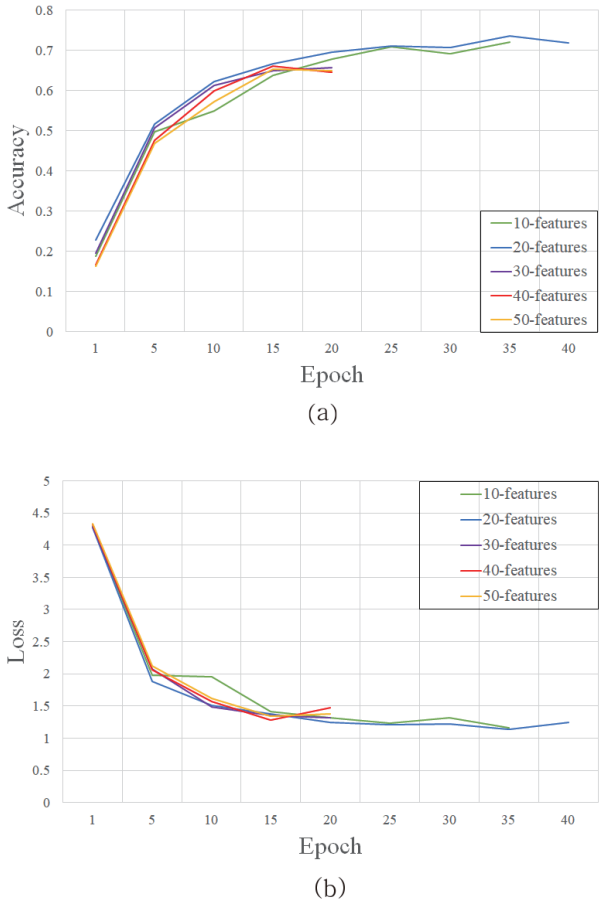


그림 6. 프레임 길이를 세분화하여 UCF-101 데이터 셋으로 학습한 CNN-LSTM 모델의 행동 인식 정확도를 측정한 실험 결과
(a) 평가 데이터 셋으로 측정한 행동 인식 모델의 정확도
(b) 평가 데이터 셋으로 측정한 행동 인식 모델의 손실 값
Fig. 6. Experimental results measuring action recognition accuracy and loss value of CNN-LSTM model learned with UCF-101 dataset by subdividing sequence length
(a) Accuracy of action recognition model measured by evaluation dataset
(b) Loss values of action recognition models measured by evaluation dataset

본 논문에서 프레임 길이를 40에 맞추어 모델 학습을 진행하였고 총 5개의 클래스(자전거 1, 2, 농구 1, 2, 벤치 프레스)를 실험하였다. 실험 영상은 YouTube에서 직접 검색하여 다운로드 받았고, 다운로드 받는 기준은 4장 3절에서 언급한 것과 동일하다. 실험 결과는 표 2 (단위 %)와 같다. 표 2의 결과를 보면 원본보다 관심 영역을 추출했을 경우의 정확도가 높아진 것을 볼 수 있다. 표에서 정의한 ROI Extracted summary는 사람 객체가

인식된 경우에만 프레임을 저장하는 방식으로 대부분의 360 영상에는 시작 부분이나, 장면 전환 같은 경우 사람 객체가 존재하지 않는다. 그렇기 때문에 행동 인식 또한 되지 않는다. 이 문제점을 제거하면서 영상 요약의 효과도 얻을 수 있었다.

표 2. 객체 인식을 통한 관심 영역 추출 및 영상 요약 기법을 이용한 행동 인식 결과
Table 2. Action recognition results using object recognition applied region of interest extraction and video summarization technique

Sequence length	Class	Original	Original with CLAHE	ROI Extracted-all frames	ROI Extracted- summary
Seq 40	Biking 1	46	44	62	76
	Biking 2	19	14	80	87
	Basketball 1	12	3	15	15
	Basketball 2	7	10	20	26
	Bench press	14	9	25	38

영상을 요약하였기 때문에 해당 영상의 특징이 더욱 부각될 수 있었고 행동 인식의 정확도 또한 단순히 관심 영역을 추출하는 것에 비하여 6%에서 최대 14%가량 증가하였다. 앞서 언급한 CLAHE를 적용하였을 경우에는 오히려 결과가 좋지 않게 나왔다. 360 영상 크기가 상당히 큰데, 모델 입력 사이즈는 상당히 작으므로 리사이징을 하는 부분에서 화질에 관한 것이 상당히 손실이 발생하여 오히려 부정적인 효과를 보인 것으로 보인다. Basketball 1에서 ROI Extracted all frames와 ROI Extracted summary의 결과가 동일한 이유는 모든 프레임에 사람 객체가 추출되었기 때문에 영상 요약이 되지 않았다. Biking 2에서 원본에 비하여 61% ~ 68%가 증가하였는데 이는 Biking 2의 영상 특성 때문으로 추측된다. Biking 2의 영상은 사람이 1명 나오고, 객체 크기가 영상 크기에 비해 작다. 따라서 제거되는 배경 영역이 크기 때문에 성능 향상이 높게 이루어진다.

V. 결 론

본 논문은 360 영상을 이용하여 관심 영역을 객체 인식 기법으로 추출한 후 행동 인식에 대한 정확도를 높이는 2-Way 방법을 제안한다. 구현 및 실험 진행 결과, 이전 방식에 의한 실제 데이터는 상당히 잡음이 많으며 여러가지 효과(카메라의 움직임, 조명 등)가 존재하기 때문에 행동 인식을 제대로 해낼 수 없다. 따라서 제안한 방법을 통하여 영상을 요약하고, 관심 영역 추출 기법을 이용하여 해당 영상에 관한 특징을 추출한다. 관심 영역 추출과 함께 영상을 요약하는 기법이 관심 영역을 추출한 전체 프레임에 대한 행동 인식보다 최대14%가 증가하였고, 원본에 대한 행동 인식 보다 최대 68%가 증가하였다. 결과적으로 본 논문이 제안하는 기법은360도 영상에 대한 기존의 전처리 없이 진행한 기법과 비교하여, 높은 행동 인식 정확도를 얻을 수 있게 한다.

Acknowledgment

본 연구는 한국전력공사의 2016년 선정 기초연구개발과제 연구비에 의해 지원되었음 (과제번호: R17XA05-68), "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 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] Byoung Chul Ko. Video-based Action Recognition Research Trends. The Institute of Electronics and Information Engineers (IEIE), 44(8), pp.16-22. 2017.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Bobick, A. F., James W. D. "The recognition of human movement using temporal templates." IEEE Transactions on Pattern Analysis & Machine Intelligence 3 pp.257-267. 2001.

[17] Chen, H. S., Chen, H. T., Chen, Y. W., & Lee, S. Y. Human action recognition using star skeleton. In Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks. ACM. pp. 171-178. 2006.

[18] Weinland, D., Boyer, E., Ronfard, R. "Action recognition from arbitrary views using 3d exemplars." 2007.

[19] Efros, A. A., Berg, A. C., Mori, G., Malik, J. "Recognizing action at a distance." In null IEEE. pp. 726. 2003.

[20] Ali, S., Shah, M. "Human action recognition in videos using kinematic features and multiple instance learning." IEEE transactions on pattern analysis and machine intelligence, 32(2), pp.288-303. 2008.

[21] Wang, L., Qiao, Y., Tang, X. "Action recognition with trajectory-pooled deep-convolutional descriptors." In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4305-4314. 2015.

[22] Simonyan, K., Zisserman, A. "Two-stream convolutional networks for action recognition in videos." In Advances in neural information processing systems. pp. 568-576. 2014.

[23] Du, Y., Wang, W., & Wang, L. "Hierarchical recurrent neural network for skeleton based action recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1110-1118. 2015.

[24] Li, C., Wang, P., Wang, S., Hou, Y., & Li, W. Skeleton-based action recognition using LSTM and CNN. In 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE. pp. 585-590. 2017.

[25] Soo-Yeun S., Joo-Heon C. Human Action Recognition System Using Multi-Mode Sensor and LSTM-based Deep Learning. Transactions of the Korean Society of Mechanical Engineers A, 42(2), pp.111-121. 2018.

[26] Janghak C., Jeongmin S., Sang-il C. "Analysis of Action Recognition Performance According to Depth of Deep Neural Network." Korean Institute of Information Scientists and Engineers (KIISE), pp.1827-1829. 2018.

[27] Sang-Jo K., Shao-Heng K., Eui-Young C. "Improved the action recognition performance of hierarchical RNNs through reinforcement learning." Korea Society of Computer Information. 26(2), pp. 360-363. 2018.

[28] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Rabinovich, A. "Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1-9. 2015.

[29] He, K., Zhang, X., Ren, S., Sun, J. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778. 2016.

[30] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. "Learning spatiotemporal features with 3d convolutional networks." In Proceedings of the IEEE international conference on computer vision. pp. 4489-4497. 2015.

[31] Y-T. K. "Contrast enhancement using brightness preserving bi-histogram equalization," IEEE Transactions on Consumer Electronics, Vol.43, pp.1-8, 1997.

[32] Chen S., Ramli A. R. "Contrast Enhancement using Recursive Mean-Separate Histogram Equalization for Scalable Brightness Preservation," IEEE Transactions on Consumer Electronics, Vol.49, pp.1301-1309, 2003.

[33] Y. Wang, Q. Chen and B. M. Zhang, "Image Enhancement based on Equal Area Dualistic sub-Image Histogram Equalization Method," IEEE Transaction on Consumer Electronics, Vol.45, pp.68-75, 1999.

[34] K. S. Sim, C. P. Tso and Y. Y. Tan, "Recursive sub-image histogram equalization applied to gray scale images," Pattern Recognition Letters, Vol.28, pp.1209-1221, 2007.

[35] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization. Graphics Gems IV," Academic Press Professional, Inc., pp.474-485, 1994.

[36] VaFRIC (Variable Frame-Rate Imperial College) Dataset, https://www.doc.ic.ac.uk/~ahanda/VaFRIC/index.html (accessed Jul. 1, 2019)

약 력

박은수

2019년 가천대학교 컴퓨터공학과 학사
2019년 가천대학교 컴퓨터공학과 석사과정
2019년~현재 성균관대학교 컴퓨터교육과 석사과정
관심분야: 멀티미디어 통신 및 시스템,
딥러닝 적용 멀티미디어

유재성

2014년~현재 가천대학교 컴퓨터공학과 학사과정
주관심분야: 멀티미디어 통신 및 시스템,
딥러닝 적용 멀티미디어

류은석

1999년 고려대학교 컴퓨터학과 학사
2001년 고려대학교 컴퓨터학과 석사
2008년 고려대학교 컴퓨터학과 박사
2008년 고려대학교 연구교수
2008년~2010년 조지아공대 박사후과정
2011년~2014년 InterDigital Labs Staff Engineer
2014년~2015년 삼성전자 수석연구원/파트장
2015년~2019년 가천대학교 컴퓨터공학과 조교수
2019년~현재 성균관대학교 컴퓨터교육과 조교수
관심분야: 멀티미디어 통신 및 시스템,
비디오 코딩 및 국제 표준, HMD/VR 응용분야