일반논문 (Regular Paper)
방송공학회논문지 제25권 제3호, 2020년 5월 (JBE Vol. 25, No. 3, May 2020)
https://doi.org/10.5909/JBE.2020.25.3.1
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

다중 객체가 존재하는 ERP 영상에서 행동 인식 모델 성능 향상을 위한 전처리 기법

박 은 수[a)], 김 승 환 [a)], 류 은 석[a)‡]

Preprocessing Technique for Improving Action Recognition Performance in ERP Video with Multiple Objects

Eun-Soo Park[a)], Seunghwan Kim[a)], Eun-Seok Ryu[a)‡]

요 약

본 논문에서 Equirectangular Projection(ERP) 영상으로 행동 인식을 할 때의 문제점들을 해결할 수 있는 전처리 기법을 제안한다. 본 논문에서 제안하는 전처리 기법은 사람 객체를 행동의 주체 즉, Object of Interest(OOI)로 가정하고, OOI의 주변 영역을 ROI로 가정한다. 전처리 기법은 3개의 모듈로 이루어져 있다. I) 객체 인식 모델로 영상 내 사람 객체를 인식한다. II) 입력 영상에서 saliency map을 생성한다. III) 인식된 사람 객체와 saliency map을 이용하여 행동의 주체를 선정한다. 이후 행동 인식 모델에 선정된 행동의 주체 boundary box를 입력하여 행동 인식 성능을 높인다. 제안하는 전처리기법을 사용한 데이터를 행동 인식 모델에 입력한 방법의 성능과 원본 ERP 영상을 입력한 방법의 성능을 비교하였을 때 최대 99.6%의 성능 향상을 보이며, OOI가 감지되는 프레임만을 추출하였을 때 행동 관련 영상 요약의 효과도 볼 수 있다.

Abstract

In this paper, we propose a preprocessing technique to solve the problems of action recognition with Equirectangular Projection (ERP) video. The preprocessing technique proposed in this paper assumes the person object as the subject of action, that is, the Object of Interest (OOI), and the surrounding area of the OOI as the ROI. The preprocessing technique consists of three modules. I) Recognize person object in the image with object recognition model. II) Create a saliency map from the input image. III) Select subject of action using recognized person object and saliency map. The subject boundary box of the selected action is input to the action recognition model in order to improve the action recognition performance. When comparing the performance of the proposed preprocessing method to the action recognition model and the performance of the original ERP image input method, the performance is improved up to 99.6%, and the action is obtained when only the OOI is detected. It can also see the effects of related video summaries.

Keyword : Action recognition, Equirectangular projection, Preprocessing

Ⅰ. 서 론

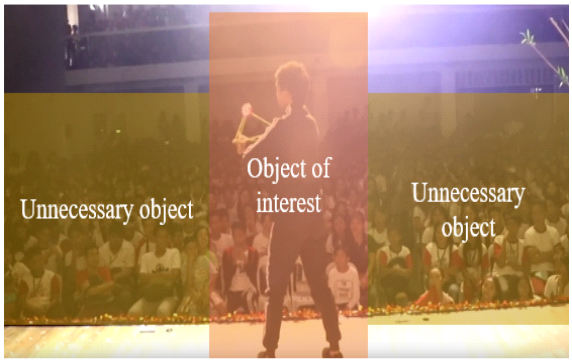
최근 고성능 GPU의 사용으로 처리 가능한 연산량이 대폭 증가함에 따라, 많은 연산이 필요한 딥 러닝 관련 기술들이 연구가 활발히 진행되어 오고 있다. 이미지 처리에 특화된 Convolution Neural Networks(CNN)의 발달로 인하여 객체 인식, 행동 인식, 이미지 캡셔닝 등과 같은 이미지 프로세싱 연구가 진행되어 오고 있다. 이 중에서 행동 인식 관련 연구는 여러 방향을 통하여 활발히 진행되어 오고 있으며, 상당히 어려운 주제로 알려져 있다.

행동 인식은 영상 데이터를 입력으로 하므로 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![1], Sports-360[2]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. 360-비디오를 재생하는 장비에서 사용자가 이질감을 느끼지 않을 정도의 재생 속도를 제공하려면 낮은 지연 속도와 Ultra-high-definition(UHD) 이상의 초고화질 360 영상이 요구된다. 이와 같은 요구사항을 해결하기 위하여 수많은 연구가 진행되어오고 있는데, 서론에서 몇 가지를 소개하도록 한다. 비대칭 코어 프로세싱 기반 타일 분할 및 할당 시스템[3,4,5]. 타일 기반 Motion-constrained tile set(MCTS)[6,7,8]. 카메라의 위치에 따른 우선순위를 적용하여 비균등 다운 샘플링을 적용한 대역폭 절감 연구[9,10]. 영상의 프로젝션 포맷 변경을 통한 대역폭 절감 연구. 이때 프로젝션 포맷은 360 영상을 2D 상에서 나타내기 위한 기법으로 Equirectangular projection(ERP), Cubemap projection(CMP) 등이 있다.

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

그림 1과 같이 ERP의 특성상 행동 인식을 할 때 불필요한 부분이 많아 성능 저하가 발생하는 것을 문제로 삼는다. 본 논문에서 ERP 영상에서 행동 인식 성능 향상을 위하여 영상 내의 행동의 주체를 찾는 방법으로 행동 인식의 성능이 향상되는 것을 보인다.

(a)



(b)

그림 1. ERP영상에서 행동 인식 시 문제점 (a) 영역적인 측면 (b) 객체적 측면
Fig. 1. Problems in Recognizing Action in ERP Image (a): Spatial Aspect (b): Object Aspect

a) 성균관대학교 컴퓨터교육과(Department of Computer Education, Sungkyunkwan University)
‡ Corresponding Author : 류은석(Eun-Seok Ryu)
E-mail:esryu@skku.edu
Tel: +82-2-760-0677
ORCID: https://orcid.org/0000-0003-4894-6105
※ 본 연구는 한국전력공사의 2016년 선정 기초연구개발과제 연구비에 의해 지원되었음 (과제번호 : R17XA05-68), "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음" (IITP-2020-2017-0-01630)(This research was supported by Korea Electric Power Corporation. (Grant number : R17XA05-68)
ITRC: "This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-2017-0-01630) supervised by the IITP(Institute for Information & communications Technology Promotion)"
· Manuscript received December 26, 2019; Revised March 4, 2020; Accepted April 16, 2020.

본 논문이 주로 기여하는 바는 다음과 같다. (1) 본 방법은 전처리 기법이기 때문에 앞으로의 행동 인식 관련 연구에서 산출되는 state of the arts 모델들을 그대로 적용 가능 (2) ERP와 같은 projection에 존재하는 왜곡을 ROI 추출 기법을 통하여 모델 입력 데이터에 포함하지 않는 것으로 행동 인식 성능 향상이 가능 (3) 행동 인식은 사람을 중심으로 하는 분야이기 때문에 사람이 없는 프레임을 제거한다면 영상의 요약이 가능 (4) 2D 이미지 데이터 셋으로 행동 인식이 가능 (5) 다중 객체가 존재하는 경우에도 영상 내의 주체가 되는 객체를 찾고 행동 인식 가능

II. 관련 연구

2장에서 본 논문에서 이용한 기술들에 관련된 연구를 소개한다. 1절에서 Saliency map과 관련된 연구를 소개하고, 2절에서 딥러닝 기술을 적용한 행동 인식 연구를 소개한다. 3절에서 360 영상과 관련된 딥러닝 연구에 관한 연구를 소개한다.

1. Saliency map

Saliency map은 영상에서 눈에 띄는 영역을 인간과 유사한 시각적 방식으로 찾아주는 것이다. 생물학적으로 인간은 색상의 차이가 심한 영역, 밝기의 차이가 심한 영역 그리고 윤곽선의 특징이 강한 영역을 먼저 본다. 이러한 특성을 기본으로 과거부터 연구가 진행되어 오고 있다[11].

전통적인 saliency 예측 방법은 contrast, rarity 그리고 symmetry 등의 특징들을 통합하는 특징 통합이론에 중점을 둔다[12,13,14]. 또 다른 연구로는 전통적인 연구들과 다르게 스펙트럼 도메인 분석을 사용한 연구가 있었다[15,16,17,18].

딥 러닝 (DL) 기반 방법[19,20,21,22,23]은 엔드 투 엔드 방식으로 saliency 예측의 정확도를 크게 향상하기 위해 어텐션 기법을 적용하였다.

2. 행동 인식

인공지능을 적용한 행동 인식과 관련된 연구를 소개하고, 그중에서도 본 논문에서 초점을 맞추고 있는 ROI 기반에 관한 연구를 소개한다.

2.1. 인공지능을 적용한 행동 인식 연구

현재의 행동 인식 연구들은 고성능 GPU를 사용한 머신러닝 기법을 사용하여 인간의 행동을 분석하는 방법이다. 머신러닝 기법을 이용한 연구들은 기존의 Rule-base 기법들 보다 대체로 성능이 높게 나오기 때문에 머신 러닝을 사용한 연구가 많이 이루어지고 있다. 그에 따른 연구로는 객체의 시간에 따른 유사도를 나타내는 Fisher vector를 잘 추출한다는 이점을 가진 Trajectory extraction 기법, 그리고 자동으로 이미지의 특징을 추출해주는 깊은 신경망 중 하나인 CNN의 장점을 취하여 성능을 높이는 연구가 있다[24]. 입력 데이터 타입을 공간 정보를 담은 이미지 데이터, 시간 정보를 담은 광학 흐름 데이터로 나누어 두 방향의 CNN을 이용한다. 그 결과를 fusion 하여 행동 인식을 하는 연구[25]. 많은 행동 인식 관련 연구들이 참고 문헌[25]을 바탕으로 진행되었다. 사람 영상을 스켈레톤 데이터로 변환하여 몸통(머리 포함), 팔, 다리 각 두 파트씩 총 5파트로 나눈 후 각각의 파트를 입력 데이터로 갖는 개량된 Recurrent Neural Network (RNN)을 사용하여 행동 인식하는 연구[26]. 스켈레톤으로 이루어진 데이터 셋을 이용하여 시퀀스별 거리 위치 특징 데이터는 Long Short-Term Memory(LSTM)에 입력한다. 공동 거리 맵은 데이터 타입이 이미지이기 때문에 CNN에 입력하는 연구[27]. 직접 센서 측정 방법으로 인간의 행동별 자료를 수집, 분석하고, 인공지능과 접목하여 인공지능 기반의 방법론에 대한 문제점들을 보완한 연구[28]. 심층 신경망의 깊이에 따른 성능 분석[29]. 강화학습을 이용하여 순환신경망의 파라미터를 최적화한 후 스켈레톤 데이터를 이용하여 행동 인식을 하는 연구[30]가 있다.

2.2. ROI 기반 행동 인식 연구

180도 카메라를 탁자 위에 놓고 전방위에 있는 “사람” 객체들을 각각 crop 하여 식습관을 판단하는 시스템에 관한 연구[31], puppet model을 이용한 연구[32], trajectory들의 움직임을 판단하여 움직임이 많은 범위를 ROI로 판단하여 행동 인식하는 연구[33], 잡음 및 ghosting 문제를 해결하기 위한 연구[34], background subtraction algorithm을 사용하여

ROI를 추출한 후 ROI 내에서 optical flow를 이용한 연구[35], attention 기반 연구[36,37], 정적 배경을 갖는 비디오에서 상호작용을 ROI 기반으로 인식하는 연구[38], saliency prediction을 통하여 ROI를 선정 후 모델에 입력할 때 해당 ROI만 입력하는 연구[39]. 요약하자면, 행동 인식 연구들은 대부분 스켈레톤 데이터를 이용하거나, 2D 카메라로 촬영한 영상 내 사람 객체가 1개일 경우 또는 2개일 경우의 상호 작용을 중점으로 연구를 진행하였다.

3. Spherical convolution

Spherical convolution 연구는 세 가지로 분류된다. i) 360 영상을 ERP로 변환 후 CNN에 입력하는 방법, ii) 360 영상에서 부분적으로 왜곡을 완화하기 위해 탄젠트 연산을 한 후 입력하는 방법, iii) i 방법은 속도는 빠르나 정확도가 낮다는 단점이 있고, ii) 방법은 그 반대이다. 기존 방법들의 단점들을 보완하기 위하여 360 영상에서 경선에 따라 중점에서부터 각도를 측정하여 각도별로 커널을 변형하는 방법을 사용하였다[40,41]. 위 3가지 방법들은 360 영상에서 객체 인식을 하기 위한 방법으로 아직 360 영상에서 행동 인식에 관한 연구는 찾아보기 힘들었다.

Ⅲ. Preprocessing Technique for Improving Action Recognition Performance in ERP Video with Multiple Objects (PIAEM)

기존의 360 영상에서 관심 영역을 추출하기 위한 연구로 Deep 360[2]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. Deep 360에서 한계

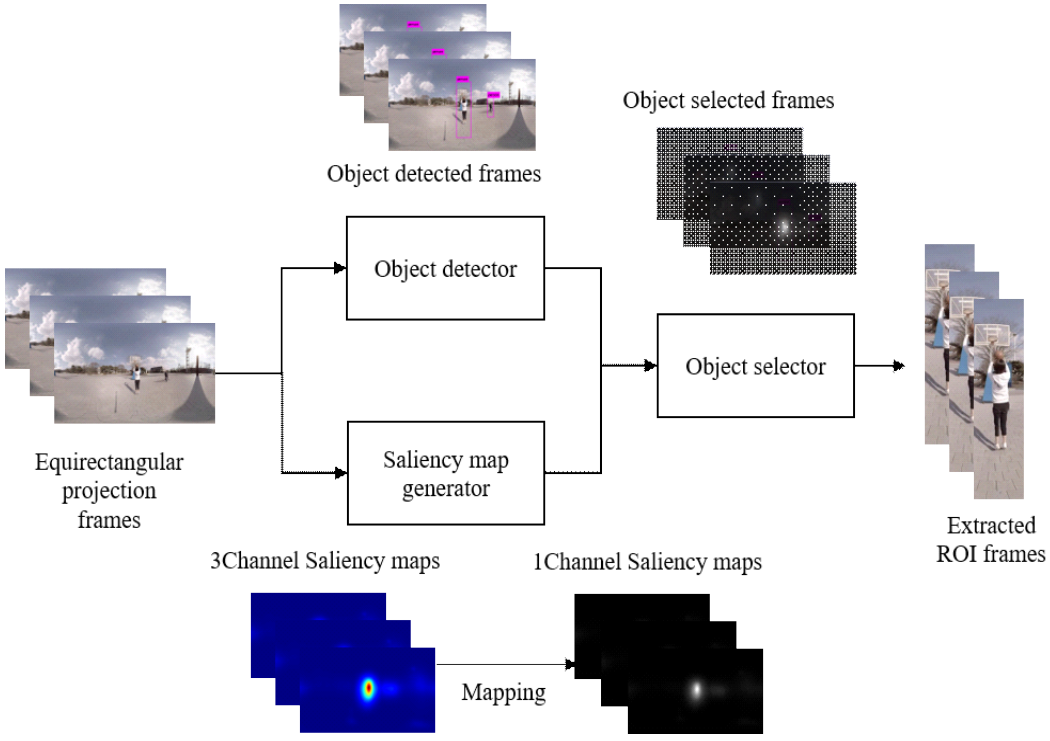


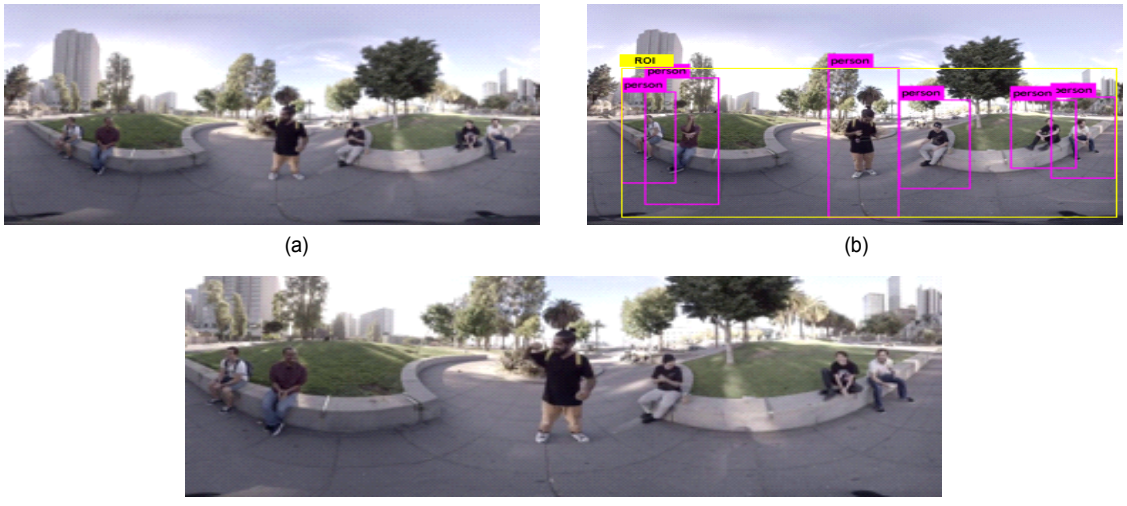
그림 2. 제안하는 ERP 영상에서 행동 인식을 할 때 성능 향상을 위한 객체인식 및 saliency map을 활용한 ROI 추출 전처리 기법의 구조도
Fig. 2. Proposed structure diagram of ROI extraction preprocessing technique using object recognition and saliency map for performance improvement in ERP imagemap for performance improvement in ERP image

점은 클래스 별로 가중치가 필요하다는 점이다. 클래스는 총 5가지 (농구, 파쿠르, 자전거, 스케이트보드, 춤)를 이용하였다. Deep 360의 한계점을 돌파하고자 행동 인식의 주체가 되는 인간 객체를 Object of interest(OOI)로 가정하고, 그 주변 영역을 ROI로 가정하였다.

2장 2절에서 언급한 것과 같이 ROI 기반 행동 인식들은 프레임 내의 인간 객체가 1개이거나, 2개의 상호작용을 중점으로 연구를 진행하였고, 이용한 데이터들의 타입은 2D 카메라로 촬영한 영상들이다. 그러나 본 논문에서 서술하는 내용은 그림 2와 같이 다중 객체가 존재하는 ERP 영상에서 객체 인식을 활용하여 관심 영역을 추출한 후 행동 인식의 성능을 향상하는 전처리 기법에 관한 연구를 진행하였다. 전처리 기법은 object detector, saliency map generator, object selector 총 3개의 모듈로 이루어져 있고, 각 모듈을 이용하여 영상 내의 주체가 되는 객체를 판단하고 영상 내의 ROI를 추출한 후 행동 인식 모델이 입력하여 행동 인식 모델 성능을 높인다.

(a) (b)

그림 3. 제안하는 전처리 기법을 위하여 YOLO 소스 수정 작업 (a) 수정 전 YOLO 결과 (b) 수정 후 결과
Fig. 3. Modification of YOLO source for proposed preprocessing technique (a) YOLO result before modification (b) Result after modification



(a) (b)

(c)

그림 4. ROI selector로 ROI 선정 과정: (a) 원본 이미지 (b) 객체 인식 모델로 모든 사람 객체를 인식한 후 ROI를 선정하는 이미지 (c) 선정된 ROI를 crop한 이미지
Fig. 4. ROI selection process with ROI selector (a) Original ERP image (b) Recognize all 사람 objects with object classification model and select ROI (c) Image cropping selected ROI

1. Object detector

본 논문에서 객체 인식 모델은 You only look once (YOLOv3)[42]를 이용하였다. YOLOv3 모델은 속도가 빠르며 다른 state-of-the-arts 모델들과 비교하여도 정확도가 뒤처지지 않는다는 장점이 있다. 본 논문에서 행동 인식의 OOI는 사람으로 가정하였으므로, 그림 3과 같이 객체 인식 모델에서 사람만을 검출하도록 한다. False acceptance rate (FAR) 을 줄이기 위하여 정확도가 90% 이상일 경우에 인식하도록 한다. ROI는 OOI의 주변 영역으로 가정하였기 때문에 기존 YOLO의 경계 박스보다 2배 크기를 늘렸다.

2. ROI selector

영상의 프레임 안에서 사람 객체는 다중, 단일, 없을 수 있다. ROI selector는 이 3가지를 고려하여 ROI를 선정한다. i) 단일일 경우: 객체 인식으로 검출된 객체의 경계 박스를 출력한다. ii) 없을 경우: 입력된 프레임을 그대로 출력한다. iii) 다중 객체일 경우 그림 4의 b와 같이 모든 사람 객체를 인식한 후 그림 4의 c와 같이 인식된 모든 객체의 경계 좌표를 이용하여 최대 영역 ROI를 선정한다. 선정하는 방법은 다음과 같다. 다중 객체의 좌표를 배열에 넣어 저장한다. 영상의 (0, 0) 좌표가 좌상단에 있기 때문에 저장된 배열 내의 값 중 가장 큰 값을 저장하는 것은 객체 경계 박스 좌표의 오른쪽 좌표, 하단 좌표이다. 가장 작은 값을 저장하는 것은 좌측 좌표, 상단 좌표이다. 이후 산출된 최대 영역 ROI 좌표들을 이용하여 영상을 crop 한다.

그러나 그림 4와 같은 방법으로 ROI를 선정할 경우 문제점이 존재한다. 행위의 주체가 아닌 객체도 ROI를 선정하는 데 영향을 끼친다는 것이다. 그림 4의 b를 볼 때 “yoyo”를 하지 않는 관중들까지 ROI 선정에 영향을 끼치는 것으로 볼 수 있다. 따라서 본 논문에서 새로운 object selector를 고안하였다.

3. Saliency map generator

본 논문에서 사용한 saliency map generator는 SAM 모델이다. SAM 모델은 Saliency Attentive Model로 Center Bias 학습이 적용된 모델이다. ERP 영상 특성상 극점으로 갈수록 왜곡이 심하여 중요도가 적어진다. 이를 위하여 적도지역에 가중치를 주는 것을 Center Bias 옵션이라고 한다[43]. SAM의 출력값은 RGB 값이 모두 있는 3 채널 saliency map으로 연산을 쉽게 하기 위하여 1 채널로 픽셀 값을 매핑한다. 그림 5와 같이 SAM의 출력값인 saliency map을 채널 변환을 할 때 일반적인 rgb2gray 방식으론 제대로 된 결괏값이 나오지 않는다. 그 이유는 SAM의 출력값인 saliency map은 JET 컬러맵이 적용되어 있기 때문이다. 따라서 JET 컬러맵과 그레이 스케일 컬러 맵(0~255)을 매핑하였다.

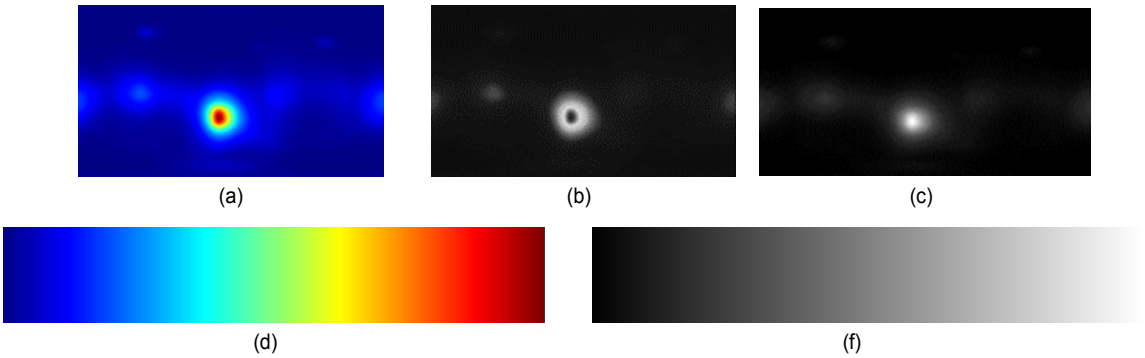


그림 5. Saliency map 전처리 과정 (a) SAM 결과 saliency map (b) RGB to Grayscale 적용 결과 (c) JET to Grayscale 적용 결과 (d) JET 컬러 모델 색상 범위 (e) Grayscale 컬러 모델 색상 범위

Fig. 5. Saliency map preprocessing (a) Saliency map output from SAM (b) RGB to GRAY result (c) JET to GRAY result (d) JET color model (e) Grayscale color model

4. Object selector

ROI selector의 문제점을 해결하기 위해 고안한 것으로 영상 내의 행위의 주체를 찾기 위한 모듈이다. 본 논문에서 촬영된 영상의 특성상 행위의 주체는 가장 눈에 띄는 위치에 있을 것이라는 가정을 하였고, 행위의 주체를 찾기 위해 saliency map을 이용하였다. Object selector의 작동 과정은 다음과 같다:

Step 1) Object selector는 saliency map generator의 출력값인 saliency map과 객체 인식된 이미지를 입력받는다.

Step 2) 수식 1과 같이 전체 프레임에서 saliency 세기의 합을 구한 뒤 전체 픽셀 개수로 평균을 구한다.

$$X = \frac{\sum_{i=1}^{n} s_i}{N} \tag{1}$$

이때 x는 전체 프레임에서 saliency 세기의 평균값, s는 각 픽셀의 saliency 세기, N은 전체 프레임 픽셀 개수이다.

Step 3) 수식 2와 같이 인식된 객체들 각각의 경계 내부의 saliency 강도를 측정하여 경계 내부 픽셀 개수로 평균을 구한다.

$$x_j = \frac{\sum_{i=1}^{n} s_i}{n} \tag{2}$$

이때 x는 인식된 객체의 saliency 세기의 평균값, j는 객체의 index, n은 인식된 객체의 경계 박스 내부 픽셀 개수이다.

Step 4) 수식 3과 같이 Step 2에서 구한 값 을 임계값으로 하고, Step 3에서 구한 각 객체들의 값 x을 임계값과 비교하여, 임계값 이상인 객체를 OOI로 선정한다.

$$OOI_k = \begin{cases} x_j, & x_j < X \\ \varnothing, & x_j \leq X \end{cases} \tag{3}$$

Step 5) 선정된 OOI는 경계 box와 함께 crop하고, 행동 인식 모델에 입력된다.

이 과정에서 얻을 수 있는 이점은 다중 객체에서 영상의 주체가 되는 객체만을 추출할 수 있는 것이다. 또한 ROI 추출로 인하여 행동 인식에 불필요한 영역을 제거할 수 있다.

Ⅳ. 실험 및 분석

4 섹션에서 본 논문에서 제안한 내용에 관한 실험의 결과를 보인다. 1절에서 본 논문의 실험 환경에서 이용한 데이터 셋에 관한 설명을 한다. 2절에서 실험 환경에 관해 자세하게 설명을 한다. 3절에서 YOLO를 ERP 영상에 사용한 이유에 관한 실험을 정성적 평가로 설명한다. 4절에서 전처리하지 않은 프레임과 ROI를 추출한 프레임을 CNN에 입력하였을 때 포커싱되는 부분을 시각화하여 표현하고, 분석한다. 5절에서 전처리 적용 후의 성능 평가를 진행한다.

1. 데이터 셋

본 논문에서 제안하는 전처리 기법에 사용되는 행동 인식 모델은 ERP 영상의 행동 인식을 하지만, 2D 이미지 데이터로 학습을 진행한다. 1절에서 학습 데이터를 서술하고, 2절에서 테스트 데이터를 서술한다.

1.1 학습 데이터 셋

행동 인식 모델 학습에 사용된 데이터 셋은 행동 인식에 많이 사용되는 UCF-101을 사용하였다. UCF-101 데이터 셋은 University of Central Florida에서 제작한 데이터 셋이다[44]. YouTube에서 다운받은 101가지 인간의 행동에 관련된 데이터 셋이 포함되어 있다. 총 13,320개의 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.

1.2 테스트 데이터 셋

본 논문에서 실험에 이용한 데이터 셋은 Youtube에서 다

운로드받은 일반적인 ERP 영상이다. 데이터셋의 선정 기준은 1) 행동 인식 모델을 UCF-101 데이터 셋으로 학습하였기 때문에 UCF-101 데이터 셋의 클래스에 해당하는 영상 데이터로 선정하였다. 2) 오브젝트의 개수는 관계없이 역동적으로 움직이는 데이터 셋으로 선정하였다.

2. Implementation Details

본 논문에서의 실험 환경은 Ubuntu 18.04 운영체제, Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz, NVIDIA GTX 1080Ti 그래픽 카드를 탑재한 PC를 이용하였다. 모듈별 소요시간을 다음과 같이 서술한다(측정 시간은 모두 평균 소요시간이다.):

객체를 인식하고 각 객체의 경계 박스의 좌표를 추출하는 과정에 걸리는 시간은 이미지의 특성에 따라 다르지만 6s가 소요된다. saliency map을 1 channel로 변환할 때 픽셀당 4.7ms가 소요된다. Saliency map과 원본 프레임의 값을 비교하여 OOI를 추출하는 과정은 300ms가 소요된다.

행동 인식을 위해 본 논문에서 사용한 모델은 CNN-LSTM이다. CNN은 Inception v3를 이용하였다. UCF-101 데이터 셋으로 학습 후 테스트했을 때 정확도는 70%이다.

3. ERP영상에서 YOLO 모델을 이용한 객체 인식 성능 실험

본 논문에서 제안하는 전처리 기법에서 객체 인식 모델을 YOLO로 선정을 하였다. 이는 그림 6과 같은 실험 결과로 보았을 때 YOLO가 ERP영상에서도 객체 인식의 성능이 높은 것을 확인하였기 때문이다. 그림 6의 “yoyo2”과 “golfswing”에 사람객체의 끊어짐 현상이 발생하였음에도 불구하고 객체를 인식하는 것이 보인다. 또한 ERP영상 속 상당히 작은 객체도 감지하는 것을 보인다. 이처럼 YOLO가 ERP 영상에서도 높은 성능을 보이는 이유는 객체 인식을 할 때 영상 전체를 한 번에 탐지하지 않고 anchor box를 이용하기 때문이다. 물론 YOLO 뿐만 아닌 Faster-RCNN과 같은 모델도 마찬가지이지만, 제안하는 전처리의 전체적 러닝타임을 낮추기 위하여 YOLO로 검증을 하였다.

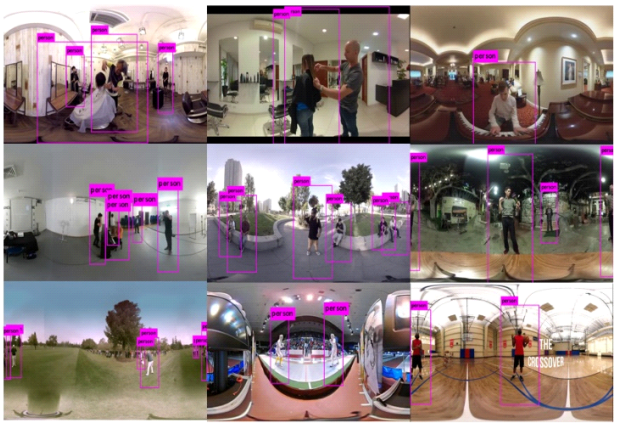


그림 6. ERP 영상에서 YOLO 모델의 객체 인식 성능 실험: 좌상단부터 우측으로 “haircut1”, “haircut2”, “piano1”, “piano2”, “yoyo1”, “yoyo2”, “golfswing”, “fencing”, “basketball”이다.

Fig. 6. Object recognition performance experiment of YOLO model in ERP image: From top left to right are “haircut1”, “haircut2”, “piano1”, “piano2”, “yoyo1”, “yoyo2”, “golfswing”, “fencing”, and “basketball”.

4. 시각화

본 논문에서 ROI를 추출하여 행동 인식의 성능을 높이는 것을 제안하였다. 제안한 방법을 이용하였을 때의 효과에 대하여 시각적으로 표현하여 이해를 돕는다. 그림 7은 히트맵 이미지로, CNN에 이미지를 입력할 경우 행동 인식을

(a)

(b)

그림 7. CNN모델에 이미지를 입력했을 때 이미지의 어떤 부분에 초점이 집중되는지 시각화한 이미지 (a) 원본 ERP 이미지를 입력 (b) ROI를 추출한 후 CNN에 입력

Fig. 7. Visualization of which part of the image is focused when the image is entered into the CNN model (a) Input original ERP image into CNN (b) Extract ROI and input it into CNN

하기 위해 이미지의 어느 부분에 초점이 맞춰지는지 알 수 있게 한 것이다. 그림 7의 (a)에서 원본 프레임에서는 행동과 관련된 위치에 부각되지 않는 것을 볼 수 있다. 이는 전체 프레임의 사이즈에 비해 행동에 관련된 정보가 담겨있는 부분이 상당히 작기 때문에 특징을 추출하기 어렵다는 것을 알 수 있다. 그에 반면에 그림 7의 (b)에서 ROI를 추출한 프레임에서 원본 프레임에서 부각되지 않았던 자전거, 사람, 피아노 등이 부각된 것이 보인다. 이는 원본 프레임의 행동 인식을 하는데 불필요한 공간을 제거하여 행동 정보가 포함된 공간의 비중이 높아졌기 때문이다.

5. 행동 인식 성능 평가

본 논문에서 여러 객체가 존재하는 ERP 영상에서 행동 인식 task를 진행할 때 성능 향상을 위하여 OOI Extractor를 제안하였다. 이전 연구인 ROI Extractor와 함께 실험을 진행하였다.

5. 1 영상 요약 실험

5개의 클래스의 비디오로 비디오 요약 및 행동 인식 성능을 표1과 같이 실험하였다. 표 1에서 총 5개의 클래스를 실험하였다. 원본 비디오보다 비교적 ROI를 추출 기법을 적용한 방법의 성능이 높다. ROI Extracted summary frames 기법은 프레임 내에 사람 객체가 존재하는 프레임만을 추출하여 영상 요약하는 기법이다. 전체 비디오는 인트로 및 장면 전환이 존재하기 때문에 해당 기법을 적용하면 비디오의 행동 특징을 비교적 잘 추출할 수 있다.

표 1. OOI extractor를 이용한 영상 요약 실험 결과 (a) 원본 영상 (b) 전체 영상 실험 (c) 요약 영상 실험
Table 1. Experimental results of OOI extractor with videos and video summarization (a) Original (b) ROI Extracted total frames (c) ROI Extracted summary frames

Class	(a) Frames	(a)	(b) Frames	(b)	(c) Frames	(c)
Biking1	1000	46.3	1000	62.5	737	76.3
Biking2	1000	19	1000	80	754	87
Basketball1	5558	12	5558	15	5543	15
Basketball2	3000	7	3000	20	2134	26
Bench press	3000	14	3000	25	1220	38

5.2 ROI extractor와 비교 실험

앞서 설명한 ROI extractor와 OOI extractor의 성능을 커스텀 데이터 셋으로 표 2와 같이 행동 인식 정확도를 비교 분석하였다. 4.4.1에서 실험한 방법은 전체 비디오에서 정확도를 나눈 값이기 때문에 성능을 비교하기에 적절하지 않다. 표 2에서, 총 5개의 클래스를 실험하였다. 각 클래스당 시퀀스 길이는 16 frames이다. CNN-LSTM 학습 시 시퀀스 길이를 8로 설정하였다. 16 프레임이기 때문에, 원본 비디오로 실험하였을 때 행동을 인식하지 못하는 경우가 많았다. 그러나 ROI 및 OOI를 추출한 후 실험을 할 경우 인식률이 크게 상승한다. ROI와 OOI의 성능 차이는 사람 객체가 한 프레임 안에 다중으로 존재할 경우 일어난다. 표 2에서 horse 클래스의 성능이 OOI가 ROI보다 낮은데 이는 사람의 객체 바운딩 박스 내에서 말을 인식하기 어려워 생긴 문제이다.

표 2. OOI extractor와 ROI extractor의 성능을 비교한 실험 결과
Table 2. Experimental results of OOI extractor with videos and compare with ROI extractor

Class	Frames	Original	ROI	OOI
GolfSwing	16	0.0	4.7	17.7
Biking	16	99.8	99.9	99.9
Basketball	16	0.0	23.6	42.9
Billiards	16	0.0	87.5	99.6
horse	16	0.0	6.1	2.9

V. 결론 및 향후 연구

본 논문에서 다중 객체가 존재하는 ERP 비디오에서 행동 인식의 성능 향상을 위한 전처리 기법을 제안하였다. 4장의 실험과 같이 원본 비디오를 행동 인식하였을 때보다 성능이 최대 99.6% 상승함을 보였다. “사람” 객체가 존재하는 프레임만 추출하여 행동 인식을 진행한 결과가 ROI를 추출한 결과보다 최대 13.7%증가한 것으로 보아 비디오 내의 전체적인 행동에 관한 요약이 가능하다는 것을 증명한다. 향후 연구는 다음과 같다. ERP 영상은 360 영상을 2D로 나타낸 영상 데이터이다. 즉, 360 영상의 특정 중심점을 기준으로 픽셀단위로 2D로 매핑이 된다. 이때 의도치 않은

객체 잘림 현상이 일어날 수 있기 때문에 행동 인식 성능이 낮아질 수 있다. 따라서 ERP 영상의 또다른 문제점인 객체 잘림 문제를 해결하는 모듈을 추가하는 연구가 필요하다.

참 고 문 헌 (References)

[1] J. Gutierrez, E. J. David, A. Coutrot, M. P. Da Silva, and P. L. Callet. 2018. Introducing UN Salient360! Benchmark: A platform for evaluating visual attention models for 360? contents. In 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX). 1?3. https://doi.org/10.1109/QoMEX.2018.8463369

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis & Machine Intelligence, (11), 1254-1259.

[12] Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research, 40(10-12), 1489-1506.

[13] Itti L. Koch C. Niebur E. (1998). A model for saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 1254?1259.

[14] Parkhurst D. Law K. Niebur E. (2002). Modeling the role of salience in the allocation of overt visual attention. Vision Research, 42, 107?123.

[15] Hou, X., & Zhang, L. (2007, June). Saliency detection: A spectral residual approach. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). Ieee.

[16] Hou, X., Harel, J., & Koch, C. (2011). Image signature: Highlighting sparse salient regions. IEEE transactions on pattern analysis and machine intelligence, 34(1), 194-201.

[17] Schauerte, B., & Stiefelhagen, R. (2012, October). Quaternion-based spectral saliency detection for eye fixation prediction. In European Conference on Computer Vision (pp. 116-129). Springer, Berlin, Heidelberg.

[18] Li, J., Levine, M. D., An, X., Xu, X., & He, H. (2012). Visual saliency based on scale-space analysis in the frequency domain. IEEE transactions on pattern analysis and machine intelligence, 35(4), 996-1010.

[19] Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 262-270).

[20] Kruthiventi, S. S., Ayush, K., & Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing, 26(9), 4446-4456.

[21] Pan, J., Sayrol, E., Giro-i-Nieto, X., McGuinness, K., & O'Connor, N. E. (2016). Shallow and deep convolutional networks for saliency prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 598-606).

[22] Wang, L., Wang, L., Lu, H., Zhang, P., & Ruan, X. (2016, October). Saliency detection with recurrent fully convolutional networks. In European conference on computer vision (pp. 825-841). Springer, Cham.

[23] Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Predicting human eye fixations via an lstm-based saliency attentive model. IEEE Transactions on Image Processing, 27(10), 5142-5154.

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

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

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

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

[28] 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.

[29] 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.

[30] 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.

[31] Rouast, P. V., & Adam, M. T. (2019). Learning deep representations for video-based intake gesture detection. arXiv preprint arXiv: 1909.10695.

[32] 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).

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

[34] Arseneau, S., & Cooperstock, J. R. (1999, August). Real-time image segmentation for action recognition. In 1999 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM 1999). Conference Proceedings (Cat. No. 99CH36368) (pp. 86-89). IEEE.

[35] Niu, F., & Abdel-Mottaleb, M. (2004, December). View-invariant human activity recognition based on shape and motion features. In IEEE Sixth International Symposium on Multimedia Software Engineering (pp. 546-556). IEEE.

[36] Sudhakaran, S., Escalera, S., & Lanz, O. (2019). Lsta: Long short-term attention for egocentric action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 9954-9963).

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

[38] 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.

[39] Sydorov, V., Alahari, K., & Schmid, C. (2019, September). Focused Attention for Action Recognition.

[40] 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).

[41] Su, Y. C., & Grauman, K. (2019). Kernel transformer networks for compact spherical convolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 9442-9451).

[42] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[43] Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). SAM: Pushing the Limits of Saliency Prediction Models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 1890-1892).

[44] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.

저 자 소 개

박 은 수

- 2013년 3월 ~ 2019년 2월 : 가천대학교 컴퓨터공학과 학사
- 2019년 3월 ~ 2019년 9월 : 가천대학교 컴퓨터공학과 석사
- 2019년 9월 ~ 현재 : 성균관대학교 컴퓨터교육과 석사
- ORCID : https://orcid.org/0000-0003-2474-3776
- 주관심분야 : 멀티미디어 통신 및 시스템, 인공지능

김 승 환

- 2014년 3월 ~ 2020년 2월 : 가천대학교 컴퓨터공학과 학사
- 2020년 3월 ~ 현재 : 성균관대학교 컴퓨터교육과 석사
- ORCID : https://orcid.org/0000-0002-7018-5114
- 주관심분야 : 멀티미디어 통신 및 시스템, 인공지능

저 자 소 개

류 은 석

- 1999년 2월 : 고려대학교 컴퓨터학과 학사
- 2001년 2월 : 고려대학교 컴퓨터학과 석사
- 2008년 2월 : 고려대학교 컴퓨터학과 박사
- 2008년 3월 ~ 2008년 8월 : 고려대학교 연구교수
- 2008년 9월 ~ 2010년 12월 : 조지아공대 박사후과정
- 2011년 1월 ~ 2014년 2월 : InterDigital Labs Staff Engineer
- 2014년 3월 ~ 2015년 2월 : 삼성전자 수석연구원/파트장
- 2015년 3월 ~ 2019년 9월 : 가천대학교 컴퓨터공학과 조교수
- 2019년 9월 ~ 현재 : 성균관대학교 컴퓨터교육학과 조교수
- ORCID : https://orcid.org/0000-0003-4894-6105
- 주관심분야 : 멀티미디어 통신 및 시스템, 비디오 코딩 및 국제 표준, HMD/VR 응용 분야