

Skeleton Keypoints를 활용한 CNN3D 기반의 버스 승객 승하차 예측모델

(CNN3D-Based Bus Passenger Prediction Model Using Skeleton Keypoints)

장진*, 김수형**

(Jin Jang, Soo Hyung Kim)

요약

버스는 대중적으로 많이 이용되는 교통수단이다. 그만큼 승객의 안전관리를 위해 철저한 대비가 필요하다. 하지만 2018년 승차하기 위해 접근하는 노인을 인지하지 못하고 버스가 출발하면서 사망사고가 발생하는 등 안전 시스템이 미흡한 상황이다. 기존에 뒷문 계단 쪽 센서를 통해 끼임 사고를 방지하는 안전 시스템은 있지만, 이러한 시스템은 위 사고처럼 승하차하려는 과정에서 발생하는 사고를 예방하진 못한다. 버스 승객의 승하차 의도를 예측할 수 있다면, 위와 같은 사고를 예방하는 안전 시스템 개발에 도움이 될 것이다. 그러나 승객의 승하차 의도를 예측하는 연구는 부족한 상태이다. 따라서 본 논문에서는 버스에 부착된 카메라 영상에서 UDP-Pose를 통해 승객의 skeleton keypoints를 추출하고, 이를 활용한 1x1 CNN3D 기반의 버스 승객 승하차 의도를 예측하는 모델을 제안한다. 제안한 모델은 승객의 승하차 의도를 예측하는 부분에서 RNN, LSTM 모델보다 약 1~2% 높은 정확도를 보여준다.

■ 중심어 : 자세 추정 ; 행동 인식

Abstract

Buses are a popular means of transportation. As such, thorough preparation is needed for passenger safety management. However, the safety system is insufficient because there are accidents such as a death accident occurred when the bus departed without recognizing the elderly approaching to get on in 2018. There is a safety system that prevents pinching accidents through sensors on the back door stairs, but such a system does not prevent accidents that occur in the process of getting on and off like the above accident. If it is possible to predict the intention of bus passengers to get on and off, it will help to develop a safety system to prevent such accidents. However, studies predicting the intention of passengers to get on and off are insufficient. Therefore, in this paper, we propose a 1x1 CNN3D-based getting on and off intention prediction model using skeleton keypoints of passengers extracted from the camera image attached to the bus through UDP-Pose. The proposed model shows approximately 1~2% higher accuracy than the RNN and LSTM models in predicting passenger's getting on and off intentions.

■ keywords : Pose estimation ; Action recognition

I. 서 론

버스는 대중교통에서 이용률이 높은 수단 중 하나이다. 그중 시내버스는 1920년 대구광역시에서

처음 도입된 이후로 2022년 현재까지도 100년 넘게 사용되고 있는 교통수단이다. 많이 이용되는 교통수단인 만큼 승차, 하차 시에 안전사고 또한 자주 발생한다. 2018년 서울 강남구에서는 버스가 정류장에 정차한 후 출발하는 과정에서 버스를 타기 위해 접근하는 노인을 치고 지나가 숨진

* 준회원, 전남대학교 인공지능융합학과 대학원생

** 종신회원, 전남대학교 AI융합대학 인공지능학부 교수

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1I1A3A04036408)

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (NRF-2020R1A4A1019191)

사고가 있었다[1]. 버스 기사가 버스에 승차하려는 승객을 미처 파악하지 못해 발생한 사고였다.

기존 버스 안전 시스템으로는 뒷문 계단 쪽에 센서가 있어 끼임을 인지해 문이 자동으로 다시 열리는 시스템이 있지만, 위 사고처럼 승하차하려는 과정에서의 사고를 방지하는 시스템은 아니다[2]. 만약 승하차 의도를 예측해서 승하차하려는 승객이 있을 때 자동으로 브레이크를 해주는 등의 안전 시스템이 있었다면 위 사고를 예방할 수 있었을 것이다. 정류장에 부착한 카메라를 통해 승객의 승하차를 판단하는 연구도 진행되었지만, 버스 출입문을 인식하여 근처에 사람이 있는지 없는지 판단하는 판별 시스템일 뿐 버스 승객의 승하차 의도 예측과는 거리가 멀었다[3]. 그 외 버스 승객의 승하차 의도를 예측하는 연구는 미흡한 실정이다.

버스 승객의 승하차 의도를 예측할 수 있다면 위에서 언급한 사고를 예방할 수 있는 안전 시스템을 개발하는 것에 도움이 될 것이다. 따라서 본 논문에서는 skeleton keypoints를 활용한 버스 승객의 승하차 의도 예측모델을 제안한다.

승하차 의도 예측은 사람의 행동 인식(Action Recognition)이 가능해야 한다. 행동 인식은 사람의 관절 좌표인 skeleton keypoints를 추출하는 자세 추정(Pose Estimation)방식을 통해 얻은 사람의 특징을 사용하여 예측할 수 있다[4,5]. 자세 추정방식은 상향식과 하향식이 있다. 상향식은 이미지에 있는 모든 관절 좌표를 추출하고 추출된 좌표의 연관성을 분석하여 자세를 추정하는 방식으로 OpenPose[6]와 HigherHRNet[7] 등이 있다. 하향식은 이미지에서 먼저 사람을 bounding box 형태로 찾은 뒤 그 내부에서 자세를 추정하는 방식으로 AlphaPose[8]와 UDP-Pose[9] 등이 있다. 본 논문에서는 자세 추정 성능이 우수한 UDP-Pose를 활용해 승객의 skeleton keypoints를 추출하고, 1x1 CNN3D 기반의 승하차 의도 예측모델을 제안한다. 행동 인식 모델로 자주 사용되는 RNN, LSTM과 비교

하여 정확도가 약 1~2% 더 높은 모델을 설계하였고, 이를 통해 승하차 의도를 결정하는데 시간에 따른 관절 좌표의 변위가 유의미함을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 자세 추정과 행동 인식의 관련 연구를 기술한다. 3장에서는 제안한 행동 인식 모델에 대해 기술하고 4장에서는 실험을 분석하고 평가한다. 마지막으로 5장에서는 결론 및 향후 연구에 대해 기술한다.

II. 관련 연구

1. 자세 추정(Pose Estimation)

자세 추정은 사람의 신체 관절인 keypoints의 위치를 추정하는 문제이다. Cao는 관절의 관계를 벡터로 표현하는 Part Affinity Fields와 관절의 위치를 나타내는 Part Confidence Maps를 이용해 자세를 추정하는 OpenPose를 제안하였다[6]. Cheng은 고해상도의 feature map을 생성하여 크기가 작은 사람에 대해서도 자세를 추정할 수 있는 HigherHRNet을 제안하였다[7]. Fang은 SSTM(Symmetric Spatial Transformer Network), NMS(Non-Maximum Suppression), PGPG(Pose-Guided Proposals Generator) 세 가지 방법을 활용한 AlphaPose를 제안하였다[8]. Huang은 자세 추정 모델에서 편향되지 않은 좌표계 변환과 편향되지 않은 키 포인트 형식 변환으로 구성된 Unbiased Data Processing(UDP)을 사용하는 UDP-Pose를 제안하였다[9].

표 1은 위에서 언급한 자세 추정 모델의 성능을 정밀도(AP)로 비교한 것이다. COCO 데이터셋[10]을 사용하여 측정하였고 AP@0.5:0.95는 OKS(Object Keypoint Similarity)를 0.5에서 0.95까지 0.05의 간격으로 AP를 측정한 결과의 평균을 의미한다[11]. OKS는 수식 (1)과 같으며 d_i 는 검출된 keypoint와 그것과 대응되는 ground truth의 유클리드 거리, v_i 는 ground

truth의 visibility flags, s 는 객체 세그먼트 영역의 제곱근, k_i 는 keypoint마다의 상수값을 의미 한다[11]. 본 논문에서는 승객의 skeleton keypoints를 검출하는 데 있어서 비교 모델들보다 AP@0.5:0.95가 0.808로 우수한 성능을 보여주는 UDP-Pose를 활용하여 실험하였다.

표 1. Keypoints Leaderboard on COCO Test-dev 2015[12]

Method	AP@ 0.5:0.95	AP@ 0.5	AP@ 0.75	AP medium	AP large
OpenPose (CMU-Pose)	0.618	0.849	0.675	0.571	0.682
AlphaPose (Fast-20-FPS)	0.717	0.888	0.783	0.674	0.780
HigherHRNet	0.721	0.895	0.784	0.681	0.775
UDP-Pose (XForwardAI)	0.808	0.949	0.881	0.774	0.857

$$OKS = \frac{\sum_i \exp\left(\frac{-d_i^2}{2s^2 k_i^2}\right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (1)$$

2. 행동 인식(Action Recognition)

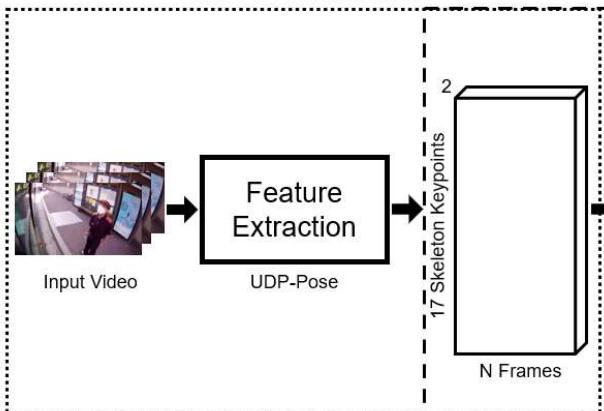
행동 인식은 사람의 특정 행동을 인식하는 연구 분야로 카메라 영상을 기반으로 사람의 관절 좌표를 추출하여 인공신경망을 통해 예측할 수 있다[13,14]. 영상에서 추출된 관절 좌표는 frame 단위로 시계열 데이터의 특성을 가진다. 그래서

사람의 행동을 인식하기 위해 인공신경망 모델 중 시계열 데이터에 대해 높은 예측 성능을 보여주는 RNN(Recurrent Neural Network)을 사용한다[13]. 시계열 데이터의 sequence가 늘어날수록 과거의 정보를 망각하는 기울기 소실 문제를 해결하기 위해 RNN의 은닉층(Hidden State)안에 기억층(Cell State)을 추가한 LSTM(Long Short-Term Memory)도 사용된다[15,16]. 위 RNN, LSTM 등의 모델을 이용해 사람의 행동 중 걷기, 서있기, 쓰러지기 등을 인식하는 연구는 진행된 사례가 있지만[13,16], 버스 승객의 승하차 의도를 예측하는 연구는 미흡했다. 이에 본 논문에서는 RNN, LSTM 모델을 이용해 승하차 의도를 예측하고 정확도를 측정했지만 정확도가 만족스럽지 않았고, 본래 영상의 특징을 추출하는 CNN3D(Convolutional Neural Network 3D)[17]를 skeleton keypoints를 활용한 승하차 예측에 사용하였다. 그 결과 RNN, LSTM보다 높은 예측 정확도를 얻었고, 시간에 따른 관절 좌표의 변위가 승하차 의도를 예측하는데 유의미함을 알 수 있었다.

III. 버스 승하차 의도 예측모델

본 논문에서는 카메라 영상에서 UDP-Pose를 활용해 각 사람의 skeleton keypoints를 x, y 좌표값으로 추출한 후 1x1 CNN3D를 기반으로 버

1. Pose Estimation



2. Action Recognition

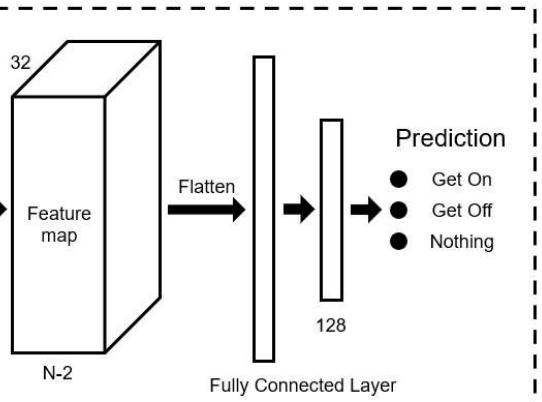


그림 1. 버스 승객의 승하차 의도 예측 과정

스 승객의 승하차 의도를 예측하는 모델을 제안 한다. 그럼 1은 본 논문에서 제안하는 버스 승객의 승하차 의도를 예측하는 과정이다. Pose Estimation 단계에서 카메라 영상은 Input Video로 사용되어 UDP-Pose를 통해 frame 단위의 x, y 좌푯값으로 변환된다. 이때 사람 객체를 추적하여 사람마다 frame 별로 keypoints 좌표 17개를 가지게 된다. 추출한 데이터를 기반으로 Action Recognition 단계를 수행하면 1x1 CNN3D 구조를 통해 feature map이 생성된다. 이를 본 논문에서 사용한 딥러닝 프레임워크인 PyTorch의 flatten 함수를 사용하여 1차원으로 변경하고, FC(Fully Connected) Layer[18]를 사용하면 승차(Get On), 하차(Get Off), 아무것도 아님(Nothing) 3가지 예측값을 얻을 수 있다.

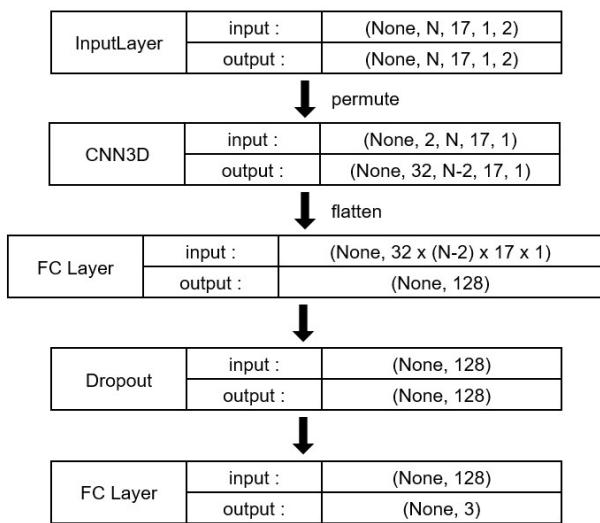


그림 2. 제안한 Action Recognition 모델

그림 2는 제안한 승하차 의도 예측모델의 구조를 입력과 출력 데이터의 형태를 포함하여 나타낸 것이다. 자세 추정 단계에서 얻은 관절 좌표를 제안한 모델에 학습하기 위해서는 관절 좌표의 데이터 차원을 변경할 필요가 있다. CNN3D 기반 모델의 학습 데이터 형태는 (B, T, H, W, C)로 5차원이며 B는 batch size, T는 time series, H는 height, W는 width, C는 channel을 의미한다[19]. HxW 크기를 갖는 이미지는 RGB와 같은 channel 수를 가지고 있고, T만큼의

frame 수를 가지면 영상이 되어 batch size 크기 만큼 학습 데이터로 사용된다[19]. 본 논문에서는 추출된 skeleton keypoints를 위와 같은 형태의 학습 데이터로 만들기 위해 다음과 같은 작업을 진행하였다. 첫 번째로 keypoints 좌푯값 x, y를 2개의 channel로 간주하였다. RNN, LSTM과 같은 행동 인식 모델은 x, y를 서로 다른 특성으로 보고 예측을 수행한다[13,16]. 그렇지만 제안한 모델에서는 관절 하나의 특성에서 파생된 x, y를 승하차 특징을 추출하는 과정에서 channel 형태로 함께 고려하였다. 두 번째로 HxW를 17x1로 대체하였다. 이미지는 가로와 세로를 갖는 2차원이지만, 추출한 skeleton keypoints는 관절 좌푯값을 일렬로 나열한 1차원이므로 dummy 차원을 생성하여 CNN3D 학습에 적합한 형태로 변경하였다. 마지막으로 time series를 의미하는 T는 승하차 의도를 예측할 때 관절 좌표를 몇 frame까지 고려할 것인지를 결정하는 매개변수로 3, 6, 9 frames로 바꾸어 가며 실험을 진행하였다. T를 N이라는 변수로 표기하면 (B, N, 17, 1, 2) 형태의 학습 데이터가 만들어진다. batch size는 실험 결과에서 다룰 것이므로 None으로 표기하여 위에서 언급한 (None, N, 17, 1, 2) 형태의 데이터가 InputLayer로 들어간다. PyTorch의 CNN3D input size는 (B, C, T, H, W)이므로[20], permute 함수를 사용하여 각 차원의 데이터 값은 유지한 채 (None, 2, N, 17, 1)로 형태를 변경해준다. 3x1x1 Convolution 3D filter 32개를 사용하면 channel이 2개에서 filter 수인 32개로 증가한다. CNN3D의 매개변수인 padding을 0으로 하여 N frames를 kernel size 3으로 Convolution 연산하면 출력은 N-2 frames로 변경된다. 그 후 flatten 함수를 사용하여 1차원으로 만들면 32x(N-2)x17x1개의 특성을 가진 데이터가 된다. 이를 FC Layer에 넣어 중간에 128차원 벡터로 변환한 뒤 승차(Get On), 하차(Get Off), 아무것도 아님(Nothing) 3가지 예측값으로 분류를 진행하였다. FC Layer 사이에는

과적합을 방지하기 위해 Dropout[21]을 추가하였다. RNN, LSTM은 데이터의 시계열을 중점으로 예측을 수행하는 모델이지만, 본 논문에서 제안하는 모델은 skeleton keypoints의 각 좌표를 단위 시간에 대해 CNN3D의 filter를 사용하여 승객의 승하차 의도를 나타내는 특징을 추출했다는 것이 차이점이다.

IV. 실험 및 결과

1. 데이터셋 설명

한국지능정보사회진흥원이 운영하는 AI 통합 플랫폼인 AI-Hub의 버스 승객 승하차 영상 데이터를 활용하였다[22]. 이 데이터셋은 대중교통 버스에 장착된 카메라로 영상 데이터를 수집해 승객마다 skeleton keypoints, 승하차 여부 등이 라벨링되어 Training, Validation 데이터로 구분되어 있다. 버스에 부착된 카메라는 총 3개로 그림 3과 같

이 A, B, C가 있다. 본 논문에서는 버스 정문에서 정류소 방향을 관측하는 카메라인 A 채널만을 사용하였다. 데이터는 2300건 내외의 영상으로 구성되어있고 하나의 영상은 평균 40초이며 3fps(frame per second)로 추출된 이미지이므로 사용된 전체 이미지는 약 27만 장이다.

모델 학습을 위해서는 영상에서 승객별로 skeleton keypoints를 추출한 데이터셋이 필요하다. 그림 4는 본 논문에서 사용한 데이터셋을 포함하여 행동 인식 모델의 학습 과정을 나타낸 것이다. 행동 인식 모델이 영상으로 승하차 의도를 예측하는 성능을 판단하기 위해 UDP-Pose를 활용하여 추출한 keypoints 17개 좌표와 라벨링된 승하차 여부로 UDP-Pose Dataset을 구축하였다. 또한, UDP-Pose와 같은 자세를 추정하는 모델 성능의 영향을 받지 않고 행동 인식 모델의 성능을 비교하기 위해 AI-Hub 데이터셋에서 라벨링 형태로 제공하는 keypoints 16개 좌표와 라벨링된 승하차 여부로 Ground Truth Dataset을 구축하였다.



그림 3. 카메라 영상 채널

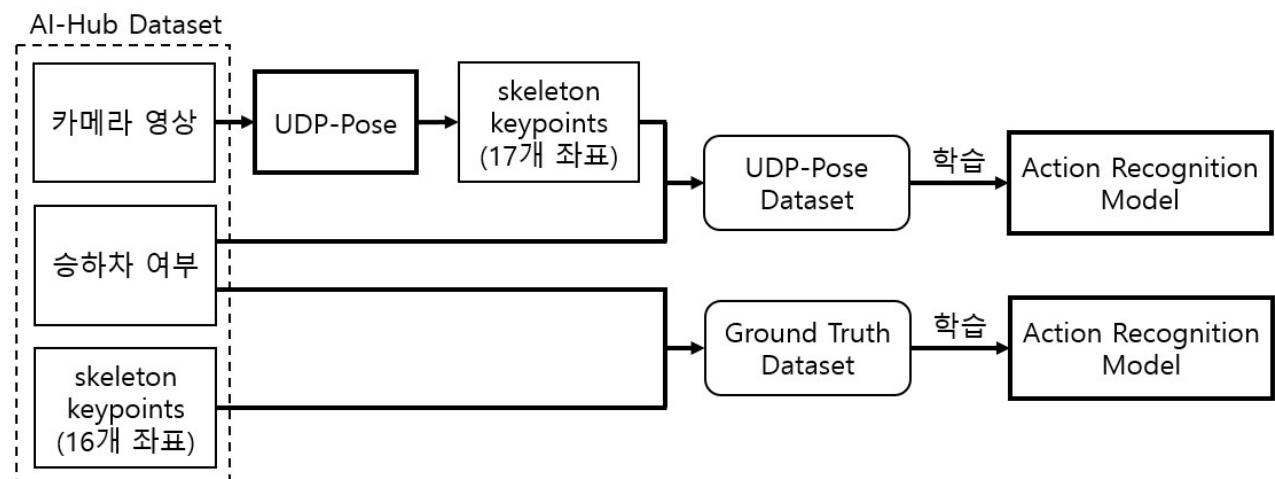


그림 4. Action Recognition 모델 학습 과정

그림 5는 두 데이터셋의 관절 좌표점을 사진에서 나타낸 것이다. UDP-Pose를 사용하여 관절 좌표를 얻은 UDP-Pose Dataset은 관련 연구에서 언급한 AP@0.5:0.95가 0.808의 성능으로 관절 좌표 17개를 추출한다[9]. 반면, AI-Hub 데이터셋의 라벨링된 관절 좌표를 사용한 Ground Truth Dataset은 사람이 직접 라벨링 하여 작성해놓은 것으로 관절 좌표 16개를 표시하고 있다[22]. UDP-Pose Dataset은 얼굴에 5개의 좌표가 집중되어 있지만, Ground Truth Dataset은 가슴과 엉덩이에 좌표점을 가지고 있다. 모델에 학습 시킬 때는 데이터셋의 종류에 따라 좌표의 개수를 나타내는 17을 16으로 바꾸어 사용하였다.



그림 5. (a) UDP-Pose Dataset, (b) Ground Truth Dataset

표 2. Train 데이터셋

Time Series	UDP-Pose Dataset	Ground Truth Dataset	Rate(%)
3 frames	79,379	138,108	57.48
6 frames	53,474	99,306	53.85
9 frames	37,805	73,831	51.20

표 3. Test 데이터셋

Time Series	UDP-Pose Dataset	Ground Truth Dataset	Rate(%)
3 frames	12,171	21,781	55.88
6 frames	7,808	14,864	52.53
9 frames	5,297	10,509	50.40

표 2와 3은 구축한 Train, Test 데이터셋의 개수를 정리한 것이다. 관절 좌표를 몇 개의 frame 까지 고려하여 예측을 수행할지 결정하는 time series를 3, 6, 9 frames로 바꾸어 가며 얻은 데이터의 개수이다. time series는 데이터가 3fps로 구성되어있어 1, 2, 3초를 의미하는 3, 6, 9 frames로 선택하였다. UDP-Pose Dataset은 Ground Truth Dataset에 비해 데이터의 수가 절반 정도의 수준으로 승객의 관절 좌표 데이터가 만들어졌다. 그 이유는 AI-Hub 데이터셋의 라벨링된 keypoints는 사람이 직접 라벨링 하여, 승객이 물체(나무, 정류장, 가로등 등)에 가려진 상황이나 버스가 이동 중일 때 인도를 걷는 승객 등의 자세 추정이 어려운 상황에도 라벨링이 되어 있기 때문이다[22]. time series가 커질수록 데이터 수가 감소하는 것은 승객의 관절 좌표가 검출되어야 하는 구간이 길어질수록 검출 오차가 커지기 때문이고, UDP-Pose Dataset의 검출 비율(Rate)이 time series가 커질수록 감소하는 이유는 라벨링된 데이터보다 UDP-Pose를 사용하여 추출한 데이터가 더 큰 검출 오차를 갖기 때문이다.

2. 실험 환경

실험에 사용한 하드웨어는 AMD Ryzen 7 3700X 8-Core Processor CPU와 NVIDIA GeForce RTX 3090 24GB GPU, 48GB RAM, Windows 10 운영체제가 설치된 데스크톱 PC이다. 프로그래밍 언어는 Python 3.6.8 버전을 사용하였고 딥러닝 프레임워크는 PyTorch 1.10.1+cu113 버전을 사용하였다.

3. 실험 결과 및 성능 평가

제안한 1x1 CNN3D 기반의 모델과 성능 비교를 위해 행동 인식에 자주 사용되는 시계열 모델인 RNN, LSTM, Attention Mechanism[23]을 포함한 LSTM과 비교하였다. 승차(Get On), 하

차(Get Off), 아무것도 아님(Nothing) 3가지를 분류하는 문제기 때문에 손실 함수(Loss Function)로 PyTorch의 CrossEntropyLoss 함수를 사용하였고, 특징 추출된 데이터의 비율이 약 15:35:50으로 아무것도 아님(Nothing)의 비중이 절반을 차지하는 불균형 데이터였기 때문에 PyTorch의 WeightedRandomSampler 함수를 사용하여 골고루 학습할 수 있게 하였다. batch size는 128로 선택하였고, 학습을 위한 Optimizer로 Adam[24]을 사용하였다. 과적합을 막고 학습 시간을 줄이기 위해 PyTorch의 EarlyStopping 함수와 Model-Checkpoint 함수를 사용하였다[25]. 평가지표는 분류 문제에 대표적으로 사용하는 정확도(Accuracy)와 각 클래스에 대하여 F1-Score를 구한 뒤 클래스의 데이터 개수에 따라 가중평균한 Weighted F1-Score를 사용하였다[26]. F1-Score는 재현율(Recall)과 정밀도(Precision)의 조화평균(Harmonic Mean)을 통해 구하며 수식은 (2)와 같다. 정확도와 Weighted F1-Score 모두 0~1 사이의 값을 가지며 1에 가까울수록 좋은 성능을 나타낸다.

$$F1 - Score = \frac{2 \times recall \times precision}{recall + precision} \quad (2)$$

표 4는 UDP-Pose Dataset에 대해 위에서 언급한 모델과 제안한 모델의 성능 평가를 비교한 것이다. UDP-Pose를 활용하여 skeleton keypoints를 추출했을 때 제안한 CNN3D를 사용한 모델이 3, 6, 9 frames에 대해 정확도 0.8223, 0.8403, 0.8566으로 가장 좋은 성능을 보여주었고

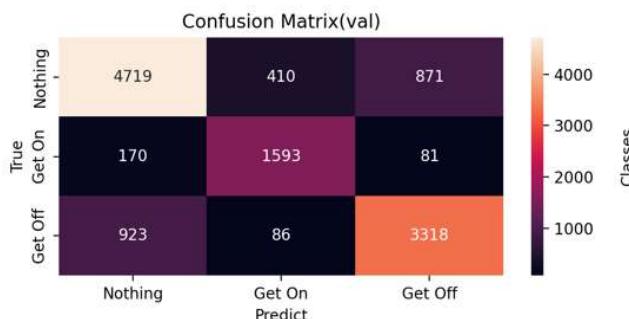


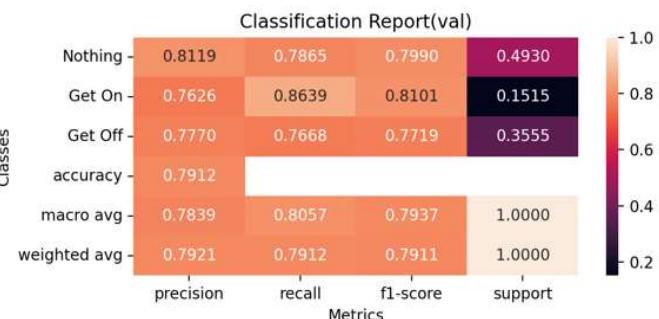
그림 6. UDP-Pose Dataset에 대한 RNN의 성능평가지표

Weighted F1-Score도 정확도와 비슷한 수치를 보여주었다.

표 4. UDP-Pose Dataset 성능 비교

Time Series	Model	정확도	Weighted F1-Score
3 frames	RNN	0.7932	0.7911
	LSTM	0.7972	0.7953
	LSTM (Attention)	0.8126	0.8107
	CNN3D	0.8223	0.8208
6 frames	RNN	0.8307	0.8313
	LSTM	0.8256	0.8257
	LSTM (Attention)	0.8390	0.8394
	CNN3D	0.8403	0.8409
9 frames	RNN	0.8400	0.8382
	LSTM	0.8369	0.8342
	LSTM (Attention)	0.8452	0.8438
	CNN3D	0.8566	0.8555

그림 6-9은 3 frames로 구성된 UDP-Pose Dataset으로 학습한 모델들의 Confusion Matrix와 정확도, 재현율, 정밀도, F1-Score를 승차(Get On), 하차(Get Off), 아무것도 아님(Nothing) 3가지 그룹에 대해서 나타낸 것이다. 제안한 CNN3D 모델이 다른 모델과 비교하여



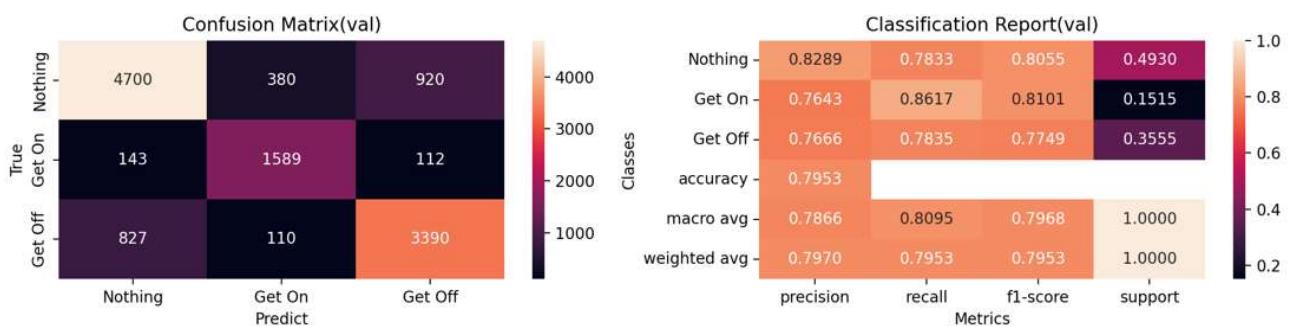


그림 7. UDP-Pose Dataset에 대한 LSTM의 성능평가지표

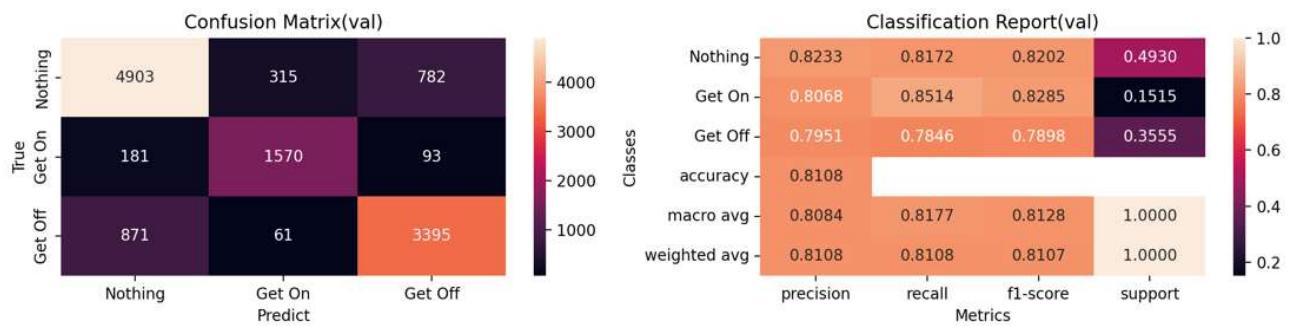


그림 8. UDP-Pose Dataset에 대한 LSTM(Attention)의 성능평가지표

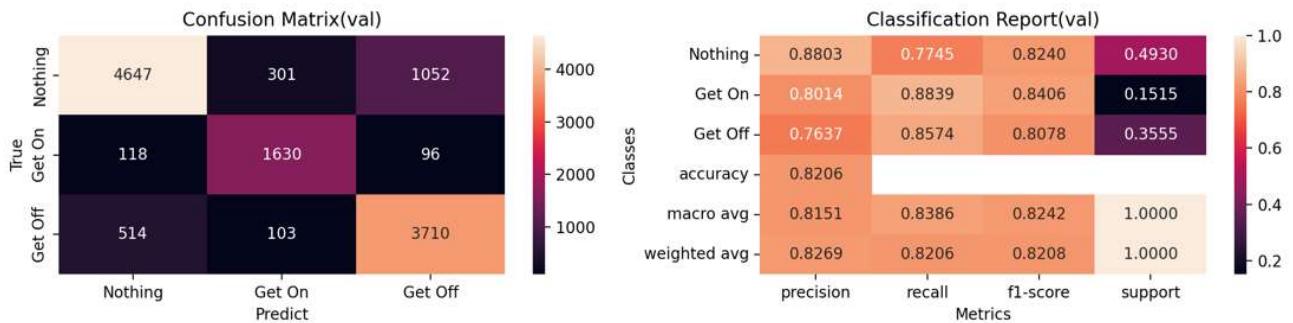


그림 9. UDP-Pose Dataset에 대한 CNN3D의 성능평가지표

F1-Score는 승차(Get On), 하차(Get Off), 아무것도 아님(Nothing)에 대해 0.8406, 0.8078, 0.8240으로 가장 좋은 성능을 보여준다. 하지만 그림 9를 보면 아무것도 아님(Nothing)의 재현율과 하차(Get Off)의 정밀도는 타 모델과 비교하여 수치가 낮은 것을 확인할 수 있다. 그 이유는 그림 9의 Confusion Matrix를 보면 아무것도 아님(Nothing)으로 예측해야 할 사람을 하차(Get Off)로 예측한 경우가 많았기 때문이다. 그렇지만 전체적인 성능을 고려했을 때 승차(Get On), 하차(Get Off)할 승객을 다른 모델보다 잘

예측했고, 재현율, 정밀도, F1-Score의 종합적인 지표인 Weighted F1-Score가 상대적으로 높으므로 의미 있는 실험 결과로 보인다.

표 5는 Ground Truth Dataset에 대해 모델 성능 평가를 비교한 것이다. time series에 대해 각각 정확도가 0.8321, 0.8512, 0.8643으로 제안한 모델이 가장 좋은 성능을 보여주었다.

그림 10-13은 3 frames로 구성된 Ground Truth Dataset으로 학습한 모델들의 Confusion Matrix와 정확도, 재현율, 정밀도, F1-Score를 승차(Get On), 하차(Get Off), 아무것도 아님

(Nothing) 3가지 그룹에 대해서 나타낸 것이다. UDP-Pose Dataset에서의 결과와 비슷하게 제안한 CNN3D 모델은 하차(Get Off)의 재현율이 LSTM(Attention) 모델보다 수치가 낮거나, 승차(Get On)의 정밀도가 LSTM 모델보다 낮은 것과 같이 다른 모델과 비교하여 모든 면에서 뛰어난 모델은 아니다. 그렇지만 승차(Get On), 하차(Get Off), 아무것도 아님(Nothing)의 3가지 분류 성능에 대해서는 정확도와 Weighted F1-Score 지표를 통해 상대적으로 우수한 성능을 보인다.

그림 14는 두 데이터셋에 대하여 실험에 사용한 행동 인식 모델의 정확도를 비교한 그래프이다. 두 데이터셋 모두 3, 6, 9 frames에 대해 제안한 CNN3D 모델이 RNN, LSTM, LSTM(Attention) 모델보다 정확도가 가장 높은 것을 확인할 수 있다.

제안한 모델을 사용하면 행동 인식에 자주 사용되는 RNN, LSTM 모델의 성능보다 승하차 의도를 예측하는 부분에서 정확도가 약 1~2% 높았다. 각 관절 좌표의 시간에 따른 변위를 Convolution 연산을 통해 특징으로 추출한 후 예측하면 전체 관절 좌표를 고려하는 시계열 예측

보다 버스 승객의 승하차 의도를 분류하는 것에 더 적합하다고 볼 수 있다.

표 5. Ground Truth Dataset 성능 비교

Time Series	Model	정확도	Weighted F1-Score
3 frames	RNN	0.8193	0.8192
	LSTM	0.8192	0.8186
	LSTM (Attention)	0.8183	0.8181
	CNN3D	0.8321	0.8319
6 frames	RNN	0.8293	0.8293
	LSTM	0.8283	0.8272
	LSTM (Attention)	0.8348	0.8335
	CNN3D	0.8512	0.8507
9 frames	RNN	0.8416	0.8443
	LSTM	0.8443	0.8427
	LSTM (Attention)	0.8481	0.8466
	CNN3D	0.8643	0.8629

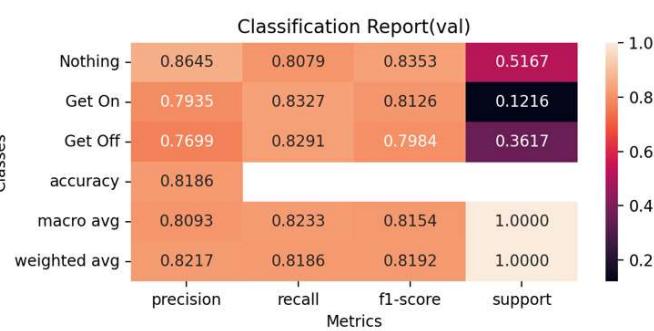
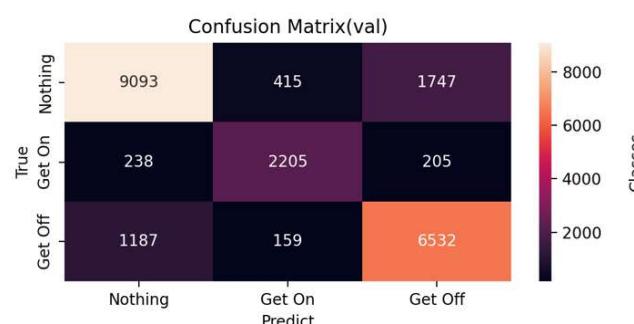


그림 10. Ground Truth Dataset에 대한 RNN의 성능평가지표

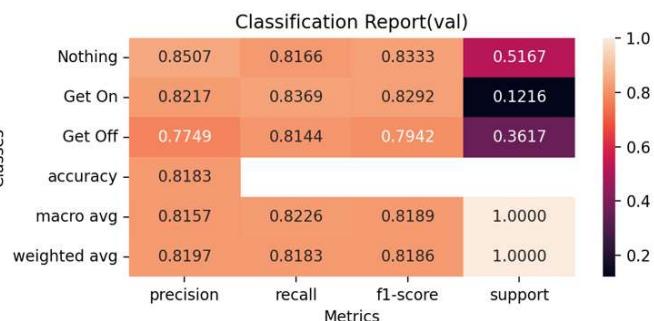


그림 11. Ground Truth Dataset에 대한 LSTM의 성능평가지표

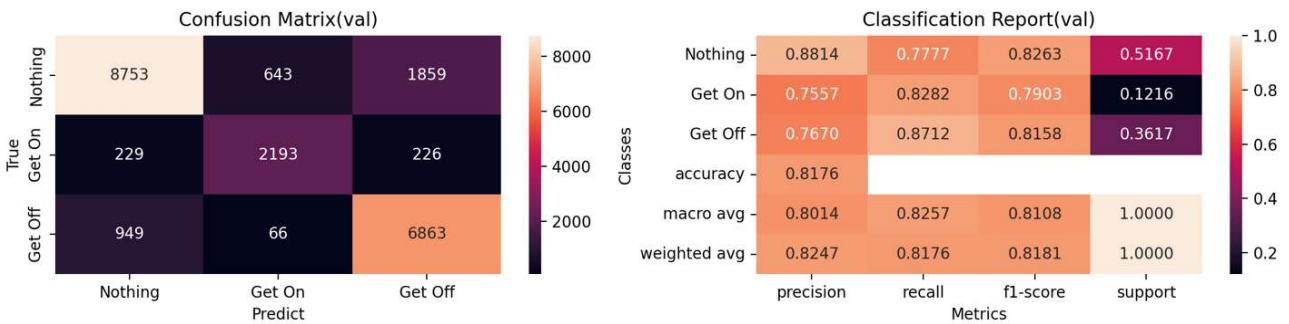


그림 12. Ground Truth Dataset에 대한 LSTM(Attention)의 성능평가지표

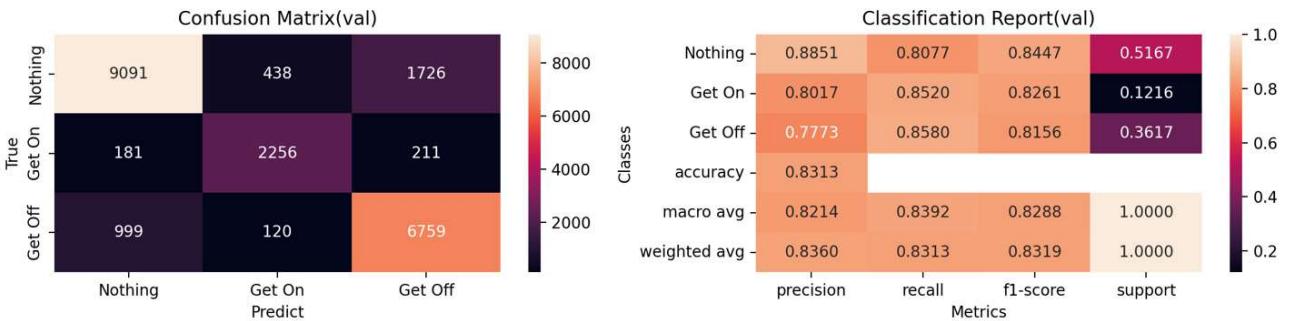


그림 13. Ground Truth Dataset에 대한 CNN3D의 성능평가지표

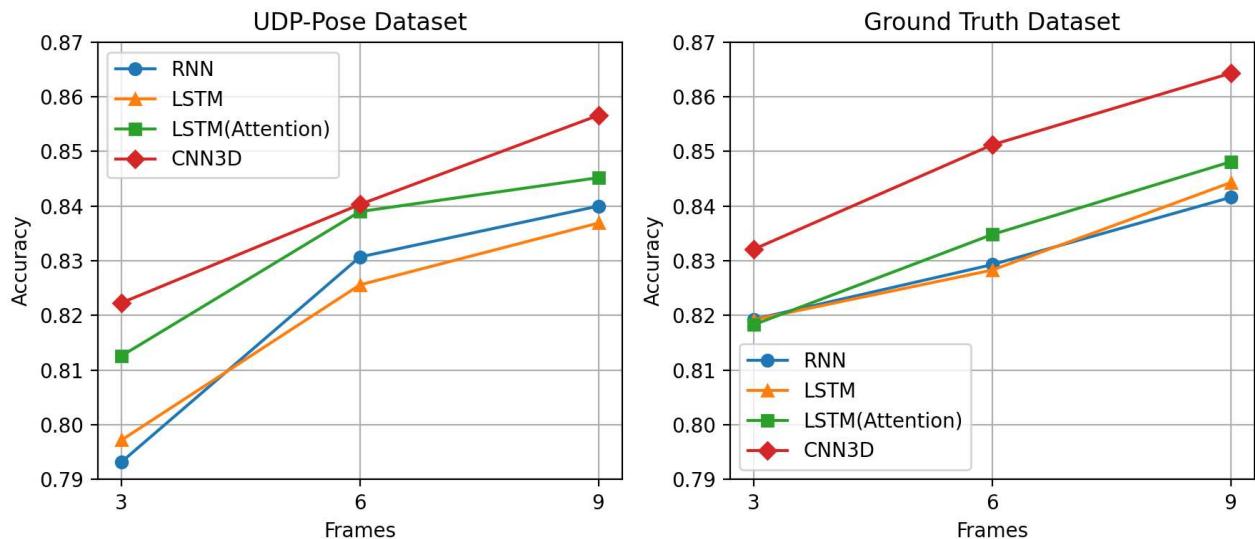


그림 14. 데이터셋에 따른 행동 인식 모델의 정확도 비교

위해 UDP-Pose Dataset과 Ground Truth Dataset을 구축하여 다른 3가지 모델과 비교하였다. 정확도와 Weighted F1-Score 평가지표를 통해 비교한 결과, RNN, LSTM, LSTM(Attention) 모델보다 제안한 CNN3D 모델이 가장 높은 성능을 보여주었다. 버스에 부착된 카메라가 1초 동안 승객을 관찰한다면 82.23%, 3초 동안 관찰한다면 85.66%의 정확도

V. 결론 및 제언

본 논문에서는 자세 추정 모델로 얻은 skeleton keypoints를 활용한 1x1 CNN3D 기반의 버스 승객 승하차 예측모델을 제안하였다. 모델 검증을

로 승객의 승하차 의도를 예측할 수 있다.

제안한 모델을 활용하여 승하차 관련 안전시스템을 개발한다면 서론에서 언급한 승하차 시 발생하는 사고 방지에 도움이 될 것으로 기대된다.

향후 연구에서는 실험 결과에서 분석했던 아무 것도 아님(Nothing)으로 예측해야 할 사람을 하차(Get Off)로 예측한 경우 등 RNN, LSTM보다 성능이 부족했던 부분을 보완하여 모든 지표에서 우수한 성능을 보여주는 모델을 설계할 계획이다. 또한, 사람의 skeleton keypoints를 x, y 값이 아닌 x, y, z의 3차원 관절 좌표를 추출하는 방법[27]을 사용하여 2차원으로는 표현되지 않는 특징을 고려한 승하차 예측 방법을 연구할 계획이다.

REFERENCES

- [1] 급히 타는 승객 못보고 출발→사망…버스기사, 벌금 형 (2 0 2 0) . https://mobile.newsis.com/view.html?ar_id=NISX20200602_0001044984 (accessed Mar., 25, 2022).
- [2] 버스 뒷문에 팔이 끼여 끌려간 여성의 사망사고! 버스는 왜!! 멈추지 않았을까? KBS 210201 방송 (2 0 2 1) . https://www.youtube.com/watch?v=C2X4_z9gUSg&t=224s (accessed Mar., 25, 2022).
- [3] 김재희, 최무룡, 윤혁진, 김대현, 조봉관, “자율주행 대중교통 시스템을 위한 정류장 승하차 판단기술,” 한국철도학회논문집, 제23권, 제4호, 339–346쪽, 2020년 4월
- [4] 박서희, 전준철, “인간 행위 인식을 위한 비전 기반 인간 자세 추정에 관한 연구,” 한국인터넷 정보학회, 제18권, 제2호, 19–25쪽, 2017년 12월
- [5] 장한별, 이칠우, “행동인식을 위한 다중 영역 기반 방사형 GCN 알고리즘,” 스마트미디어저널, 제11권, 제1호, 46–57쪽, 2022년 02월
- [6] Z. Cao, G. Hidalgo, T. Simon, S.E. Wei and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
- [7] B. Cheng, B. Xiao, J. Wang, H. Shi, T.S. Huang and L. Zhang, “HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5385–5394, 2020.
- [8] H.S. Fang, S. Xie, Y.W. Tai and C. Lu, “RMPE: Regional Multi-person Pose Estimation,” *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2353–2362, Venice, Italy, Dec. 2017.
- [9] J. Huang, Z. Zhu, F. Guo and G. Huang, “The Devil Is in the Details: Delving Into Unbiased Data Processing for Human Pose Estimation,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5699–5708, 2020.
- [10] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C.L. Zitnick, “Microsoft COCO: Common Objects in Context,” *Computer Vision - ECCV 2014*, vol. 8693, pp. 740–755, 2014.
- [11] 1. Keypoint Evaluation(2022). <https://cocodataset.org/#keypoints-eval> (accessed Mar., 25, 2022).
- [12] Keypoint Leaderboard(2022). <https://cocodataset.org/#keypoints-leaderboard> (accessed Feb., 18, 2022).
- [13] 김미경, 차의영, “스켈레톤 벡터 정보와 RNN 학습을 이용한 행동인식 알고리즘,” 방송공학회논문지, 제23권, 제5호, 598–605쪽, 2018년 9월
- [14] N.N. Hoang, G.S. Lee, S.H. Kim and H.J. Yang, “Effective Hand Gesture Recognition by Key Frame Selection and 3D Neural Network,” *Smart Media Journal*, vol. 9, no. 1, pp. 23–29, Mar. 2020.
- [15] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [16] 배현재, 장규진, 김영훈, 김진평, “AlphaPose를 활용한 LSTM(Long Short-Term Memory) 기반 이상행동인식,” 정보처리학회논문지, 제10권, 제5호, 187–194쪽, 2021년 10월
- [17] 변영현, 곽근창, “데이터 종류에 따른 딥러닝 기반 행동인식 연구동향,” 대한전기학회 학술대회 논문집, 194–195쪽, 2021년 11월
- [18] 김상조, 김미경, 차의영, “RGB 데이터 기반 행동인식에 관한 연구,” 한국정보처리학회 춘계학술발표대회 논문집, 제24권, 제1호, 936–937쪽, 2017년 4월
- [19] [Deep Learning]데이터 표현(2020). <https://someone-life.tistory.com/3> (accessed Mar., 25, 2022).
- [20] CONV3D(2019). <https://pytorch.org/docs/stable/generated/torch.nn.Conv3d.html> (accessed Mar., 25, 2022).
- [21] N. Srivastava, G. Hinton, A. Krizhevsky, I.

- Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.
- [22] 버스 승객 승하차 영상 소개(2021).
<https://aihub.or.kr/aidata/34166> (accessed Feb., 18, 2022).
- [23] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv preprint arXiv:1409.0473, 2014.
- [24] D.P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, 2014.
- [25] 김강민, 김판구, 전찬준, "Deep Metric Learning을 활용한 합성곱 신경망 기반의 피부질환 분류 기술," 스마트미디어저널, 제10권, 제4호, 45–53쪽, 2021년 12월
- [26] Classification - Metrics (2)(2021).
<https://hongl.tistory.com/136> (accessed Mar., 25, 2022).
- [27] 정근석, 박병준, 윤경로, "RGB영상과 깊이영상을 이용한 3D 휴먼 골격 키포인트 탐지," 전기학회논문지, 제70권, 제9호, 1354–1361쪽, 2021년 9월

저자 소개



장진(준회원)

2018년 한양대학교 수학과 학사 졸업.
 2019년~현재 전남대학교 인공지능융합학과 석사 과정.

<주관심분야 : 자세 추정, 행동 인식, 인공지능>



김수형(종신회원)

1986년 서울대학교 컴퓨터공학과 학사 졸업.
 1988년 KAIST 전산학과 석사 졸업.
 1993년 KAIST 전산학과 박사 졸업.
 1997년~현재 전남대학교 AI융합대학 인공지능학부 교수

<주관심분야 : 인공지능, 자연영상 패턴인식, 감정인식, 정밀의료, 문서영상처리>