The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 21, No. 6, pp.155-161, Dec. 31, 2021. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2021.21.6.155
JIIBC 2021-6-23

합성 블록 어텐션 모듈을 이용한 운동 동작 인식 성능 분석

Performance Analysis of Exercise Gesture-Recognition Using Convolutional Block Attention Module

경찬욱*, 정우용*, 선준호*, 선영규*, 김진영**

Chanuk Kyeong*, Wooyong Jung*, Joonho Seon*,
Young-Ghyu Sun*, Jin-Young Kim**

요 약 최근, 실시간으로 카메라를 통해 동작을 인식하는 기술의 연구가 많이 진행되고 있다. 기존의 연구들에서는 사람의 관절로부터 특징을 추출하는 개수가 적기 때문에 동작 분류의 정확도가 낮은 한계점들이 있다. 본 논문에서는 이러한 한계점들을 해결하기 위해 움직일 때 변하는 관절의 각도를 특징 추출하여 계산하는 알고리즘과 이미지 분류 시에 정확도가 높은 CBAM(Convolutional Block Attention Module)을 사용한 분류모델을 제안한다. AI Hub에서 제공하는 피트니스 자세 이미지로부터 5가지 운동 동작 이미지를 인용하여 분류 모델에 적용한다. 구글에서 제공하는 그래프 기반 프레임워크인 MediaPipe 기법을 사용하여, 이미지로부터 운동 동작 분류에 중요한 8가지 관절 각도 정보를 추가적으로 추출한다. 추출한 특징들을 모델의 입력으로 설정하여, 분류 모델을 학습시킨다. 시뮬레이션 결과로부터 제안한 모델은 높은 정확도로 운동 동작을 구분하는 것을 확인할 수 있다.

Abstract Gesture recognition analytics through a camera in real time have been widely studied in recent years. Since a small number of features from human joints are extracted, low accuracy of classifying models is get in conventional gesture recognition studies. In this paper, CBAM (Convolutional Block Attention Module) with high accuracy for classifying images is proposed as a classification model and algorithm calculating the angle of joints depending on actions is presented to solve the issues. Employing five exercise gestures images from the fitness posture images provided by AI Hub, the images are applied to the classification model. Important 8-joint angles information for classifying the exercise gestures is extracted from the images by using MediaPipe, a graph-based framework provided by Google. Setting the features as input of the classification model, the classification model is learned. From the simulation results, it is confirmed that the exercise gestures are classified with high accuracy in the proposed model.

Key Words : Convolutional Block Attention Module(CBAM), features-extracting, gesture-recognition, human joints, MediaPipe

*준회원, 광운대학교 전자융합공학과
**정회원, 광운대학교 전자융합공학과, 교신저자
접수일자 2021년 10월 6일, 수정완료 2021년 11월 6일
게재확정일자 2021년 12월 10일

Received: 6 October, 2021 / Revised: 6 November, 2021 / Accepted: 10 December, 2021
*Corresponding Author: jinyoung@kw.ac.kr
Dept of Wireless Communications Engineering, Kwangwoon Univ, Seoul, Korea

Ⅰ. 서 론

최근 COVID-19으로 인한 사회적 거리두기가 진행되고 있다. 그에 따라서, 운동 장소가 헬스장과 야외에서 집으로 변하고 있다. 운동장소의 변화는 비대면 홈 트레이닝 시장을 성장시킨다. 비대면 홈 트레이닝은 카메라를 통하여, 실시간으로 운동 동작을 인식하는 기술이 요구된다. 동작 인식 기술은 카메라를 통해 얻어진 신호를 분석, 처리, 가공하여 분류하는 기술로, 여러 운동 동작을 학습시켜 운동 자세 교정, 운동 종류 분석 등을 수행할 수 있다. 과거에는 키넥스 카메라를 이용하여 동작 인식을 연구하였다[1]. 하지만, 키넥스 카메라를 통한 동작 인식은 물체를 3차원으로 인식하기 위한 센서가 따로 필요하고, 동적 객체를 바탕으로 골격화를 시도하기 때문에 사람과 비슷한 객체가 존재하더라도 이를 사람이라 인식하고 골격화를 시도하고, 또한 사용자가 어떠한 물체를 들고 있는 상황에서 해당 물체까지 골격화를 시도하기 때문에 정확한 골격화 정보를 인식하지 못하는 문제가 발생한다. 최근에, 머신러닝을 통해 웹 카메라로 동작을 인식하는 기술이 활발히 연구되고 있다[2]. 머신러닝을 통한 동작 인식은 정확히 사람 몸만 인식을 하게 되고 3차원으로 임베딩이 가능하다. 또한, 과거에 사용했던 키넥트 카메라에 필요한 센서들을 사용하지 않아도 된다.

머신러닝을 이용한 동작인식 기술에서 동작을 분류할 때 DTW(Dynamic Time Warping), kNN(k-Nearest Neighbor), CNN(Convolutional Neural Network)[3], LSTM(Long-Short Term Memory)[4], GRU(Gated Recurrent Units)[5]과 RNN(Recurrent Neural Network)[6] 알고리즘을 주로 사용한다. 본 논문에서는 CBAM (Convolutional Block Attention Module) 기법을 분류모델로서 사용한다. CBAM 알고리즘은 기존의 알고리즘보다 더 높은 정확도와 빠른 연산 속도를 가진다[7].

CBAM 분류 모델을 학습시켜 정확도를 높이기 위해서는 운동 동작 시에 키 포인트가 되는 특징들을 추출하여 분류모델에 입력으로 설정해야한다. 본 논문에서는 MediaPipe으로부터 나온 skeleton 정보를 활용하여, 특징들을 추출하고 동작 인식 분류에 중요한 키 포인트 특징을 계산하여 입력으로 설정한다[8].

본 논문은 다음과 같이 구성되어있다. Ⅱ장에서는 MediaPipe를 소개하고 Ⅲ, Ⅳ장에서는 BAM과 CBAM에 대해 설명한다. Ⅴ장에서는 제안하는 시스템 모델과 특징 추출 방법에 대해 설명하고 Ⅵ장에서는 시뮬레이션 결과 제시한다. 마지막으로 Ⅶ장에서 결론으로 마무리한다.

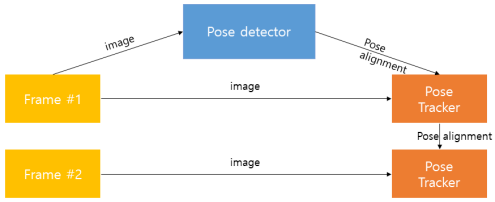

그림 1. 포즈 추정 파이프라인 개요.
Fig. 1. Pose detection pipeline summary.

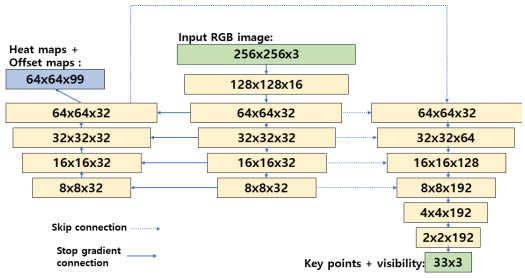

그림 2. 파이프라인의 추적 모델 구조.
Fig. 2. Detection pipeline model structure.

Ⅱ. MediaPipe

MediaPipe에서 인체 자세 인식에 대한 새로운 접근 방식인 BlazePose를 발표했다. 머신러닝을 사용하여 인간 자세 추적을 제공하여 33개의 단일 프레임에서 본 몸체의 2D 경계 부분(landmark)을 추론한다. 사용방법은 CPU추론과 GPU추론이 있다. CPU추론은 휴대폰에서 실시간 성능을 달성할 수 있고, GPU추론은 초 실시간 성능이 가능하다.

MediaPipe에서는 포즈 추정을 위해 detector-tracker ML 파이프라인을 사용한다. 파이프라인은 detector을 사용하여 프레임 내에 포즈 관심 영역(ROI)를 찾아 33개의 키 포인트를 모두 예측한다. 동영상의 경우 감지기는 첫 번째 프레임에서만 실행된다. 후속 프레임은 그림 1과 같이 이전 프레임의 포즈 키 포인트에서 ROI를 도출한다.

MediaPipe에서 파이프라인의 포즈 추정은 x, y위치 및 가시성과 3개의 가상정렬 키 포인트를 사용하여 33개의 모든 키 포인트의 위치를 예측한다. 그림 2와 같이 모든 키 포인트가 결합된 히트 맵과 오프셋 예측으로 감독되는 회귀 접근 방식을 사용한다.

훈련 중에 먼저 히트 맵과 오프셋 손실을 이용하여 네트워크의 중앙 및 왼쪽 타워를 훈련한다. 그 후, 히트 맵

출력을 제거하고 회귀 인코더인 오른쪽 타워를 훈련하므로 히트 맵을 효과적으로 사용하여 경량 임베딩을 주도한다.

Ⅲ. BAM

CNN알고리즘에서의 bottleneck은 spatial pooling이 이루어지는 부분을 말한다. spatial pooling은 CNN의 abstraction과정에서 필수적인 부분이며, 피쳐 맵의 해상도가 작아지게 된다. Bottleneck 구간에서 정보량이 줄기 전에 BAM을 추가하여, attention 모듈로 중요한 부분의 값을 키우고, 덜 중요한 부분의 값을 줄이게 된다.

BAM은 3차원의 합성곱 피쳐 맵을 입력으로 받고, attention으로 정제된 합성곱 피쳐 맵을 출력한다. 그림 4와 같이 channel축과 spatial축의 attention을 나누어 계산하고, 각 출력 값을 더하고 sigmoid함수를 통해서 입력과 같은 사이즈의 3차원 attention 맵을 생성한다. 입력 합성곱 피쳐 맵과 attention 맵을 곱해준 값을 기존 입력에 더한다[9].

channel축은 피쳐를 1차원 벡터로 만들어주는 global average pooling을 통해서 각 채널의 global context를 모아주고, 2층의 다층 퍼셉트론을 통과하여 입력 채널과 같은 사이즈를 출력한다. spatial 축은 채널이 갖는 의미를 유지하기 위해서 합성곱으로 최종 2차원 attention을 계산한다.

BAM의 디자인 방법에는 2가지 파라미터가 있는데, 채널 압축 정도(r)와 공간 attention에서의 dilation value(d)가 있다. 본 논문에서는 가장 높은 정확도를 갖는 채널 압축 정도와 dilation value를 각각 16, 4로 설정한다.

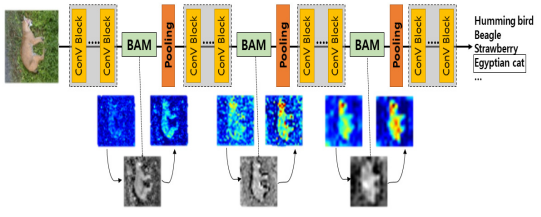

그림 3. BAM이 결합된 네트워크 구조.
Fig. 3. Network combined BAM structure.

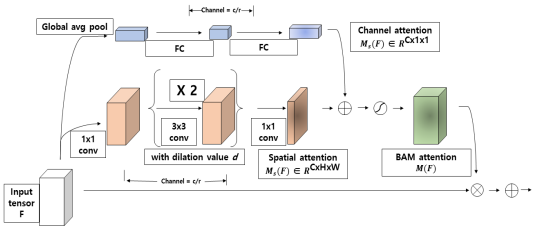

그림 4. BAM의 연산 과정.
Fig. 4. Operation process of BAM.

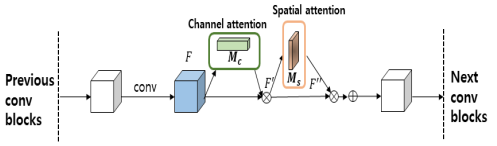

그림 5. CBAM 구조.
Fig. 5. Structure of CBAM.

Ⅳ. CBAM

CBAM은 BAM의 후속 연구로, 더욱 향상된 성능을 보여준다. 기존 BAM은 channel과 spatial 축에서 하나의 3차원으로 더해서 구현하지만, CBAM은 순차적으로 channel attention을 먼저 적용하고 spatial attention을 추가하여 더 좋은 성능을 보여준다[10].

Channel attention은, 기존 BAM은 average pooling을 사용했지만, CBAM은 average pool, max pool 두 가지를 결합하여 사용한다. pooling된 특징은 같은 의미를 공유하는 값이기 때문에, 하나의 공유된 다층 퍼셉트론을 사용할 수 있으며, 파라미터 양을 줄일 수 있다. Spatial attention 또한 대칭적으로 구성되어, 단 하나의 합성곱으로 spatial attention을 계산한다.

Ⅴ. 시스템 모델

1. AI Hub 데이터 셋

본 논문에서 모델을 학습시키기 위하여 훈련 데이터와 검증 데이터를 AI Hub의 헬스케어 피트니스 동작 데이터를 인용한다[11]. 헬스케어 피트니스 동작 데이터에서 맨몸 운동 5가지 동작을 모델에 학습시켰다. 동작은 버피 테스트, 크로스 런지, 레그레이즈, 사이드 런지, 스탠딩 사이드로 구성된다. 각 동작은 256개의 프레임으로 구성된다.

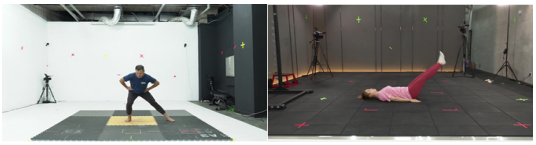

그림 6. AI Hub에서 제공하는 맨몸 운동 동작 데이터 셋.
Fig. 6. Physical exercise gesture dataset.

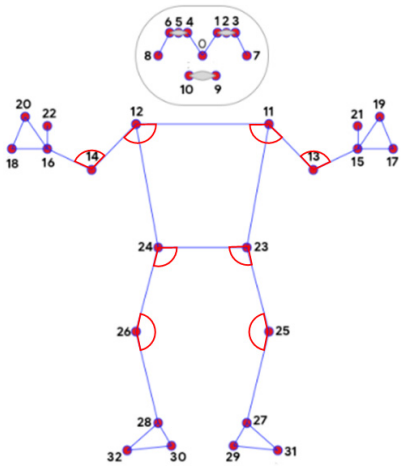

그림 7. 제안한 8개의 각도 특징 추출.
Fig. 7. Proposed 8 angel features extraction.

2. MediaPipe를 이용한 임베딩

MediaPipe는 33가지의 키 포인트를 인식하여 4가지의 특징을 추출한다. x좌표, y좌표, z좌표와 가시성으로 구성된다. 본 논문에서 MediaPipe를 통해서 얻을 수 있는 기본적인 132개의 특징과 8개의 특징을 추가하여 140개의 특징으로 데이터를 처리하여 모델을 학습시킨다. 본 논문에서 제안한 8개의 특징은 그림 7과 같다.

왼쪽 팔꿈치부터 오른쪽 무릎 순으로 각도를 순서대로 추출한다. 맨몸 운동 동작에서 가장 정보가 많이 담긴 각도이기 때문에 그림 7과 같은 8개의 관절에 대한 각도 정보를 특징으로 추출한다. 관절에 대한 각도는 식 (1)과 같이 표현된다.

$$\theta_i = \cos^{-1}\frac{v_{ij} \bullet v_{ik}}{|v_{ij}| \bullet |v_{ik}|}, \quad (1)$$

여기서, θ_i는 i번 째 각도를 가리키고, v_{ik}, v_{ij}는 θ_i를 끼인 각으로 가진 방향벡터를 의미한다. i는 각도 정보의 번호를 의미하기 때문에 14, 12, 11, 13, 24, 23, 26와 25의 값을 갖는다. j, k는 i주위에 있는 점들의 번호를 의미한 다. v_{ij}와v_{ik}는 MediaPipe에서 나온 3차원 좌표정보를 이용하여 구할 수 있다. 좌표정보를 통해 추출한v_{ij}와v_{ik}를 이용하여 식 (1)와 같이 θ_i를 구할 수 있다.

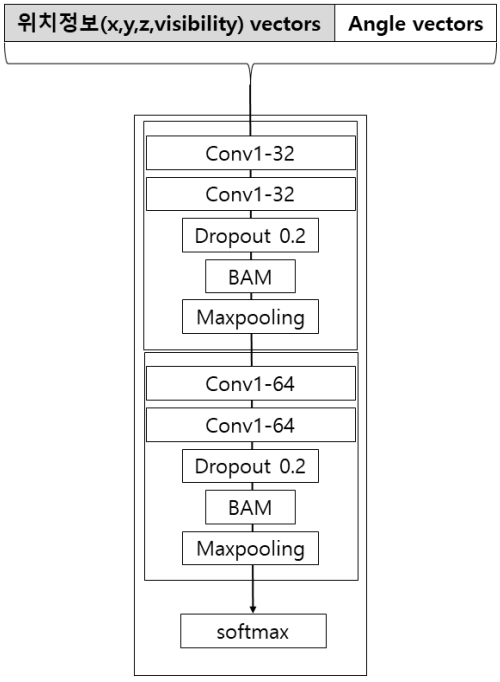

그림 8. 제안한 운동 동작 인식 모델 구조.
Fig. 8. Proposed exercise gesture recognition model structure.

3. 제안한 모델 구조

본 논문에서는 MediaPipe를 통하여 나온 132개의 특징과 중요한 관절의 각도 정보를 추가하여 140개의 특징을 추출한 후, 여러 프레임을 반복하여 훈련 데이터 셋으로 총 1080개의 프레임을 설정하고 검증 데이터 셋으로 121개의 프레임을 설정하여 학습을 진행한다. 각 프레임은 140개의 특징으로 구성된다. 그림 8과 같이 프레임에서 나온 140개의 특징은 CBAM 모델에 입력으로 설정된다. 학습된 모델은 버피 테스트, 크로스 런지, 레그 레이즈, 사이드 런지, 스탠딩 사이드와 같이 5가지 동작 중 확률이 가장 높은 동작을 출력한다[12].

Ⅵ. 실험 및 결과

본 논문에서는 5가지의 동작을 구분하는 모델을 제안한다. 그림 9와 같이, 분류모델에서 분류방법은 softmax함수를 사용하여 각 5가지 동작에 대한 확률로

결과가 나온다. 원 핫 인코딩 방식을 사용하여 실제 동작과 모델에서 나온 예측 값을 비교하여 오차를 측정한다. 오차는 범주형 교차 엔트로피 오차를 사용하여, 오차를 줄이면서 학습하는 모델을 제안한다. 오차는 식 (2)와 같이 정의된다.

$$Loss = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C} t_{ij}\log(y_{ij}), \qquad (2)$$

여기서, t는 실제 값, y는 예측 값이다. C는 동작의 개수를 표현하므로 5로 설정되고, N은 전체 프레임의 개수를 의미한다.

제안하는 모델은 epoch는 50으로 설정한다. 그림 10와 같이 epoch가 늘어날수록 오차는 감소하고 정확도는 증가하는 그래프를 확인할 수 있다.

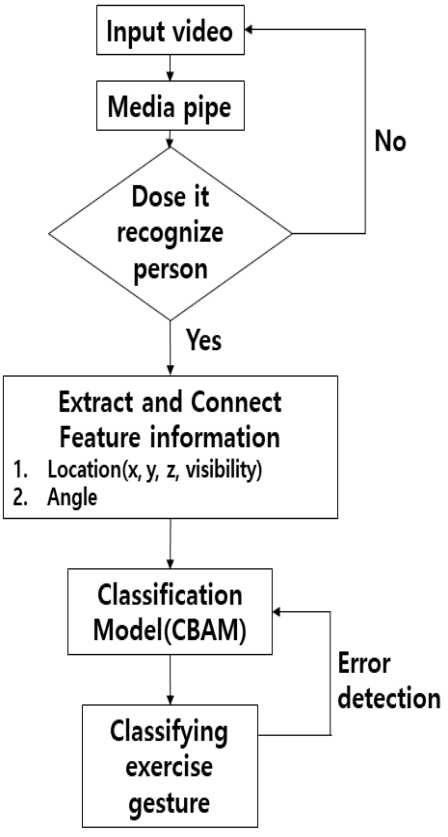

그림 9. 운동 동작 인식 알고리즘 흐름도.
Fig. 9. Exercise gesture recognition algorithm flowchart.

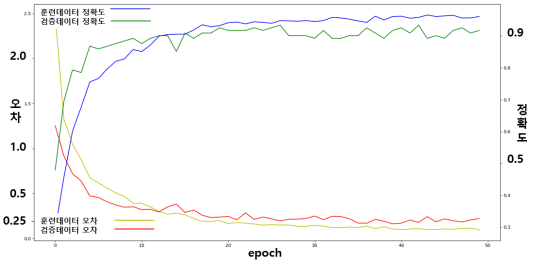

그림 10. 분류모델 정확도와 오차.
Fig. 10. Accuracy and loss of the classification model.

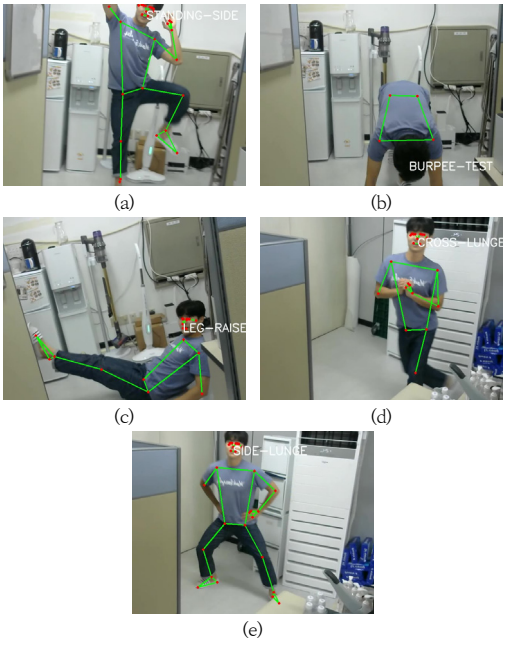

그림 11. 실시간 운동 동작 인식 결과.
Fig. 11. Results of real time exercise gesture recognition.

그림 10으로부터 검증데이터 정확도가 90% 이상 나온 것을 확인할 수 있다.

그림 11은 웹 카메라를 이용하여 제안한 알고리즘을 구현 및 동작 모습이다. 실시간으로 동작을 분류하여 각 운동 동작에 대한 이름을 카메라 모니터를 통해 보여준다. 그림 11에서 (a)는 스탠딩 사이드, (b)는 버피테스트, (c)는 레그 레이즈, (d)는 크로스 런지 그리고 (e)는 사이드 런지 로 구성된다.

VII. 결 론

본 논문에서는 MediaPipe를 이용하여 특징을 추출하여 CBAM 학습시키는 모델을 구성했다. 제안하는 방법은 MediaPipe를 통해 한 프레임에서 140개의 특징을 추출한다. 이러한 특징은 CBAM 분류모델에 입력되고 5가지의 동작에 대한 확률로 출력이 된다. 제안하는 모델은 스탠딩 사이드, 버피 테스트, 레그 레이즈, 사이드 런지와 크로스 런지를 높은 정확도로 분류한다.

본 논문에서는 분류모델에 기본적인 CNN, LSTM와 kNN 같은 방식을 사용하지 않고 더 성능이 좋은 CBAM을 사용함으로써 정확도를 높였다. 하지만, MediaPipe에서 계산한 결과로 다시 분류하는 과정에서 소요되는 시간이 길다. 성능이 좋은 GPU로 운동 동작 분류는 가능하지만, 성능이 안 좋을 경우, 분류할 때 소모되는 시간이 길다. 소모되는 시간이 길어지면, 실시간으로 운동 동작을 구분하는데 지연이 생겨, 화면이 조금씩 끊기는 현상이 발생한다. 이러한 문제점을 해결하기 위해서는 MediaPipe보다 계산량이 적은 방식을 사용할 필요가 있다.

본 논문에서 제안한 방식은 운동 동작에 국한되지 않고 다른 연구에도 활용가능하다. 이상행동 감지, 자세분석, 가상현실과 증강현실에 대한 연구에도 적용할 수 있다. 따라서, 충분한 데이터 셋을 확보하고, 데이터 셋과 모델이 적합하게 설계된다면, 다양한 연구 분야로 진출할 것으로 예상된다.

References

[1] Y. G. Sun, Y. M. Hwang, S. G. Hong and J. Y. Kim., "Performance of real-time image recognition algorithm based on machine learning," Journal of Satellite, Information and Communications, vol. 12, no. 3, pp. 69-73, Sep. 2017.
DOI:https://www.koreascience.or.kr/article/JAKO201770475988287.pdf

[2] H. S. Choi, "Kinect-based motion recognition model for the 3D contents control," The Journal of the Korea Contents Association, vol. 14, no. 1, pp. 24-29, Jan. 2014.
DOI: http://dx.doi.org/10.5392/JKCA.2014.14.01.024

[3] 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.
DOI: https://doi.org/10.1109/TPAMI.2019.2929257

[4] J. W. Park, "Egocentric activity recognition based on shot-movements aggregation using LSTM and fusion," (M. S. diss., Korea Advanced Institute of Science and Technology, Daejeon), pp. 1-24, Feb. 2018.
DOI: http://library.kaist.ac.kr/search/detail/view.do?bibCtrlNo=733776&flag=dissertation

[5] K. W. Kim, "AI-based activity recognition from early-stage motion," (M. S. diss., University of Korea, Sejong), Feb. 2020.
DOI: http://www.riss.kr/link?id=T15530467&outLink=K

[6] M. K. Kim and E. Y. Cha, "Using skeleton vector information and RNN learning behavior recognition algorithm," Journal of Broadcast Engineering, vol. 23, no. 5, Sep. 2018.
DOI: https://doi.org/10.5909/JBE.2018.23.5.598

[7] J. H. Lee, U. N. Yoon and G-S. Jo, "CNN-based speech emotion recognition model applying transfer learning and attention mechanism," Journal of Korean Institute of Information Scientists and Engineers, vol. 47, no. 7, pp. 665-673, July 2020.
DOI: https://doi.org/10.5626/JOK.2020.47.7.665

[8] G-M. Kim and J-H. Baek, "Real-time hand gesture recognition based on deep learning," Journal of Korea Multimedia Society vol. 22, no. 4, pp. 424-431, Apr. 2019.
DOI: https://doi.org/10.9717/kmms.2019.22.4.424

[9] J. Park, S. Woo, J-Y. Lee and I. S. Kweon, "BAM: Bottleneck attention module," Proceedings of the British Machine Vision Conference, Northumbria, July 2018.
DOI: https://arxiv.org/abs/1807.06514

[10] J. Park, S. Woo, J-Y. Lee and I. S. Kweon, "CBAM: Convolutional block attention module," Proceedings of the European Conference on Computer Vision, Munich, (ECCV), Germany pp. 3-19, July 2018.
DOI: https://arxiv.org/abs/1807.06521

[11] AI Hub. Available online: https://aihub.or.kr/aidata/8051.

[12] R. R. Koli and T. I. Bagban, "Human action recognition using deep neural networks," 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pp. 376-380, July 2020.
DOI: https://doi.org/10.1109/WorldS450073.2020.9210345

저 자 소 개

경 찬 욱(준회원)

- 2021년 8월 : 광운대학교 전자융합공학 학사 졸업
- 2021년 9월 ~ 현재 : 광운대학교 전자융합공학과 석박사통합과정
- 관심분야 : 인공지능, 디지털통신, 에너지 인터넷

정 우 용(준회원)

- 2017년 3월~현재 : 광운대학교 전자융합공학 학사 재학
- 관심분야 : 인공지능, 디지털통신, 영상처리

선 준 호(준회원)

- 2021년 2월 : 광운대학교 전자융합공학 학사 졸업
- 2021년 3월~현재 : 광운대학교 전자융합공학과 석박사통합과정
- 관심분야 : 인공지능, 에너지 인터넷, 디지털통신

선 영 규(준회원)

- 2018년 2월 : 광운대학교 전자융합공학 학사 졸업
- 2018년 3월~현재 : 광운대학교 전파공학과 석박사통합과정
- 관심분야 : 무선통신시스템, 딥러닝, 에너지인터넷

김 진 영(정회원)

- 1998년 2월 : 서울대학교 전자공학과 공학박사
- 2001년 2월 : SK텔레콤 네트웍 연구소 책임연구원
- 2001년 3월 ~ 현재 : 광운대학교 전자융합공학과 교수
- 관심분야 : 인공지능, 차세대이동통신, 전력선통신, 가시광통신

※ "위 논문은 문화체육관광부의 스포츠산업기술개발사업에 의거 국민체육진흥공단의 국민체육진흥기금을 지원받아 연구되었습니다. (KSPOs202101-07-03-06)"