Journal of the Korea Industrial Information Systems Research Vol.28 No.3, Jun. 2023 :1-10
http://dx.doi.org/10.9723/jksiis.2023.28.3.001 ISSN:1229-3741

영상신호를 입력으로 하는 3D ResNet기반 유아 행동 인식 기법+

(3D ResNet-based Children's Behavior Recognition Method Using Video Image Sequence)

박 재 석[1], 차 기 주[2], 최 아 영[3]*
(Jaeseok Park, Kijoo Cha, and Ahyoung Choi)

요 약 본 연구에서는 다수의 유아가 등장하는 영상 내의 행동을 인식하기 위하여 딥러닝 기반의 유아 행동 인식 기술을 개발하였다. 유아들의 경우 동일한 행동이라도 표현과 방법이 다양하여 다양한 종류의 입력에 강건하게 분석될 수 있는 딥러닝 모델에 대한 개발이 필요하다. 본 연구에서는 입력 신호를 딥러닝의 입력에 맞도록 처리하고 3D ResNet을 사용하여 행동 인식 알고리즘을 제안하였다. 50명의 유아를 대상으로 13개 행동을 수행하는 영상 자료를 수집하였으며, 실험결과 13개의 행동 인식에 평균 72.21% 정확도를 보였다. 행동 중 서 있기 90.74%, 밀고 당기기 88.89%, 앉기 90.74%의 행동 인식률을 보였다. 향후 본 연구 결과물을 통해 일상생활에서 유아들의 행동 패턴을 자동으로 분석하고 서비스하는 연구에 활용될 수 있다.

핵심주제어: 유아 행동 인식, 딥러닝, ResNet, 영상신호

Abstract In this study we propose a deep learning model to detect children behavior from the video sequence. In the case of children it is necessary to develop a deep learning model that can be robustly analyzed for various types of inputs with various expressions and methods even for the same behavior. In this study we propose an action recognition algorithm using 3D ResNet based on input sequence image. For data collection image data of performing 13 actions were acquired for 50 children and as a result of the experiment an average of 72.21% accuracy was shown in recognizing 13 actions. In the case of actions such as standing pushing and pulling and sitting the recognition rate was around 90% accuracy In the future, the results of this study can be used for research that automatically analyzes and provides services for behavior patterns of young children in daily life.

Keywords: Children behavior recognition, deep learning, ResNet, video input

* Corresponding Author: aychoi@gachon.ac.kr
+ 이 논문은 2020년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임.(2020S1A5A2A03041734)
Manuscript received April 02, 2023 / revised May 17, 2023 / accepted June 02, 2023
1) 가천대학교 산업경영공학과, 제1저자
2) 가천대학교 유아교육학과, 제2저자
3) 가천대학교 AI.소프트웨어학부, 교신저자

1. 서 론

최근 영상신호 기반으로 사용자 행동을 인식하는 연구가 활발히 수행되고 있다. 영상을 통해 연속된 이미지를 기반으로 SIFT (Scale-Invariant Feature Transform) 또는 SURF (Speeded Up Robust Features)를 사용해 이미지에서 특징을 추출하거나 실루엣, 깊이 및 골격을 인식하여 사용자의 포즈를 예측하는 연구가 있으며, 사용자의 시선을 방향을 인식하고, 이벤트와 사람의 행동을 나타내는 시나리오를 기반으로 행동 인식하는 연구도 수행되었다 (Jung et al., 2011; Shahroudy et al., 2016; Jobanputra, C. et al., 2019; Kal et al., 2019; Kim et al., 2019; Shi et al., 2019; Duan et al., 2022).

Kal et al.(2019)은 비디오 기반 사람 행동 인식을 통한 스마트 감시 시스템을 제안하였다. 시스템은 ISFT를 통해 추출한 특징을 기반으로 떨어지기, 싸우기, 걷기, 달리기, 앉기 등을 포함한 4가지에서 9가지의 행동을 인식하기 위해 K-Nearest Neighbor(KNN)와 Support Vector Machine(SVM)의 기계학습 기법을 적용했다. SVM의 결과는 92.91%인 반면 KNN의 정확도는 90.83%였다. Kim et al. (2019)는 입력 비디오로부터 관절 특징을 추출해 노인들의 행동을 인식하는 방법을 제안하였다. 히든 마르코프를 적용한 이 모델은 9일간 노인들의 일상 행동에 대하여 84.33%의 정확도를 보였다. Shahroudy et al.(2016)은 NTU RGB-D 데이터를 이용하여 2-layer Part-Aware LSTM 모델을 제안하였다. 이 모델은 40개의 일상 행동(먹기, 마시기, 읽기 등) 그리고 11개의 상호 행동(펀칭, 발차기, 안기 등)으로 구성된 60개의 클래스를 인식하는데 사용되었다. 제안된 모델은 cross subject 테스트에서 62.93%의 정확도를 보였으며, cross view 테스트에서 70.27%의 정확도를 보였다. Shi et al. (2019)는 새로운 2-스트림 적응형 그래프 컨벌루션 네트워크(2s-AGCN)를 제안하였다. 두 개의 GCN 스트림 구조를 통하여 행동 인식에 더욱 유용한 관절 데이터의 뼈의 길이와 방향 등 2차 정보를 추가로 이용할 수 있게 됨으로써 높은 유연성과 정확도를 얻을 수 있었다. 이 모델은 NTU RGB-D 데이터 기반의 실험에서 cross subject 테스트 85% cross view 테스트에서 95.1%의 성능을 보였다. Duan et al. (2022)는 골격의 표현 방식을 그래프(GCN)에서 벗어나 히트맵 스택에 의존하는 관절 기반 동작 인식 접근방식인 PoseC3d를 제안하였다. 이는 비디오의 시공간적 특징을 학습하는 데 효과적이며 추가 계산 비용이 적어 효율적이다. 이 모델은 cross subject 테스트, 즉 모델이 학습 과정에서 사용되지 않은 새로운 주체(사람 또는 개체)에 대해 얼마나 잘 작동하는지를 평가에서 94%의 성능을 보였다. 그리고 cross view 테스트, 즉 모델이 학습되지 않은 다른 시점 또는 관점에서 데이터를 평가한 결과 97.1%의 높은 성능을 보인다. 이는 모델이 새로운 주체와 다른 시점에서도 일관된 예측을 수행할 수 있는 능력을 갖추고 있음을 나타낸다.

그러나 기존 대부분의 연구는 성인을 대상으로 수집한 데이터를 기반으로 행동 인식을 주로 수행하였다. Brouwers et al. (2021)은 어린이를

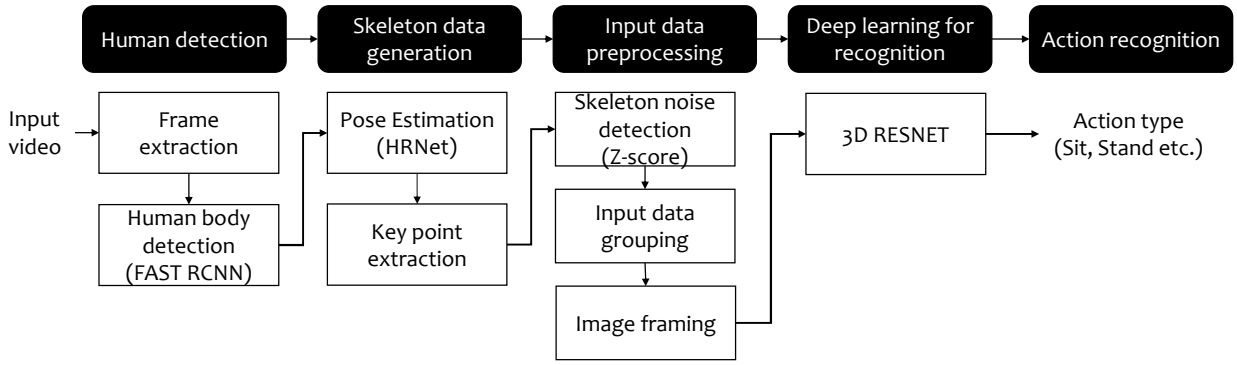

Fig. 1 Action recognition process

대상으로한 비디오 기반 스포츠 활동 인식연구를 수행하였다. 이 연구에서는 성인의 데이터로 학습된 slowfast-resnet50 기반 행동 인식이 어린이에게 적용하였을 때 네트워크 교차시 일반화된 성능이 보장되지 않음을 확인하였다. 특히 결과에서 나타난 것과 같이 21개의 행동 분석에 아동 데이터 세트로 학습된 모델을 어린이 데이터에 적용한 경우 43.8%의 정확도를 보였으며, 성인 데이터세트로 학습한 성인 모델에서 어린이 데이터를 적용한 경우 35.0%의 정확도를 보였다. 어린이 모델과 성인 모델을 혼합하여 생성한 하이브리드 모델에서는 43.8%, 46.2%의 정확도를 보였다. 낮은 정확도의 원인으로 어린이의 활동을 분석하는데 있어 더 높은 분산을 나타내기 때문이라는 결론을 얻었다. 즉, 유아들의 경우 같은 행동이라도 표현과 방법이 다양하여 다양한 종류의 입력에 강건하게 분석될 수 있는 딥러닝 모델에 대한 연구 개발이 필요하다. 또한, NTU 데이터베이스나 UCI 데이터베이스 등은 기존 연구에서 활용하는 데이터 세트는 제한된 환경에서 특정 행동을 연기하는 성인의 정면, 측면을 측정한 데이터베이스이므로 다양성이 확보되는 현장의 데이터를 기반으로 분석시 정확도가 크게 떨어질 수 있다. 따라서 본 연구에서는 정제되거나 연기하지 않은 자연스러운 다수 유아의 행동 데이터를 인식하는 모델을 개발하는 것을 목적으로 한다. 특히 유아들의 경우 같은 행동이라도 표현과 방법이 다양하여 다양한 종류의 입력에 강건하게 분석될 수 있는 딥러닝 모델에 대한 개발이 필요하다.

2. 시스템 구성 및 실험

2.1 시스템 구성

전체 시스템은 Fig. 1과 같이 구성되었다. 입력 데이터는 다수 유아의 행동이 담긴 영상 데이터를 수집하여 사용하였으며, 유아 행동으로는 Straudenmayer et al.(2009)가 제안한 유아가 놀이 활동 동안 가장 흔하게 활동하는 내용을 중심으로 행동을 구분할 수 있도록 하였다. Straudenmayer et al.(2009) 연구를 기반으로 하여 유아의 흥미 영역별로 언어영역, 수 조작영역, 역할 쌓기 영역, 음률영역, 미술영역, 과학영역, 유희실에서 하는 활동들을 유아 교육 전문가 2인 이상과 함께 선별하였다. 이후 유아에게 각 영역에서 공통적으로 관찰되는 활동인 엎드려있기, 앉기, 서있기, 걸어다니기, 달리기, 제자리 점프뛰기 등의 13개의 활동을 대표적인 유아 행동으로 정의하였다.

제안하는 시스템은 4가지 단계로 구성되어 있으며, 사용자 인식(human detection), 관절 데이터 추출(skeleton image extraction), 딥러닝 입력을 위한 잡신호 제거 및 영상 데이터 그룹화(input image data processing), 딥러닝 모델에 적용 및 결과 도출(Deep learning model for action recognition)의 순서로 행동 인식이 수행된다. 첫 번째 단계로 영상으로부터 사용자 인식을 수행하기 위해 MMCV, MMDetection 모듈을 사용하여 유아 행동 비디오의 프레임을 추출하였다(Chen et al, 2019; OpenMMLab). MMCV는 컴퓨터 비전 연구용 라이브러리이며 MMDetection은 객체 감지를 지원하는 라이브러리이다. 두 번째 단계로 사용자 인식의 결과를 바탕으로 자세 인식을 수행하여 관절 좌표를 추출하였다. 관절 좌표를 기반으로 행동인식을 수행한 이유는 관절 좌표가 영상입력을 넣는 경우에 대비하여 학습시 리소스를 최소화하고 인식 성능의 개선을 보여주었기 때문이다(Zhang et al., 2019). 사람으로 인식한 위치의 바운딩 박스에 HRNet 모델을 적용하여 17개의 관절 좌표를 추출하였다. 세 번째 단계로, 딥러닝의 모델 입력 데이터를 생성하기 위해서 노이즈 데이터를 제거 하였다. 자세 인식의 결과로 얻은 좌표를 기반으로 포즈를 생성할 때 사람의 포즈로 적용이 어려운 경우의 데이터를 제외하였다. 입력 영상에서 전신이 정면으로 보이는 경우에는 입력 영상과 동일하게 포즈 예측에 문제가 없으나, 정면을 보다가 측면을 보는 경우 인식의 시작점인 얼굴 좌표와 어깨 등의 좌표가 인식되지 않아 사람의 포즈 예측에 오류가 발생한다. 따라서 관절을 연결한 형태가 사람의 포즈로 인식이 어려운 경우를 노이즈로 판단하여 이를 제거

하였다. 이후 유아별로 입력 데이터를 생성하기 위해 정적인 행동과 동적인 행동으로 나누어 데이터를 군집화하였다. 정적인 행동의 경우 K-means clustering 알고리즘을 적용하였으며, 동적인 행동의 경우 다중 객체 인식 (Multiple object tracking, MOT) 모델을 적용하였다 (W Luo et al. 2021). 다음으로 연속된 동작을 인식하기 위하여 관절 이미지를 30프레임 단위로 묶은 후 gif 파일로 저장하였다. 프레임 양에 따라 연산량이 급격히 증가하여 본 논문에서는 [60, 60]으로 사이즈를 변환하여 저장하였다. 네 번째 단계로, 3D ResNet-18 모델을 적용하여 유아의 행동을 구분하였다.

2.2 입력 신호 전처리

딥러닝의 입력 정보를 생성하기 위해서는 사용자 인식, 관절 데이터 추출, 딥러닝 입력을 위한 잡신호 제거 및 영상 데이터 그룹화의 세 단계를 거쳤다. 첫 번째 단계로 영상으로부터 사용자 인식을 수행하기 위해 MMCV, MMDetection 모듈을 사용하여 유아 행동 비디오의 프레임을 추출하였다. 행동별 약 10초가량의 유아 행동 영상 하나당 198개의 프레임을 추출하였으며 추출한 프레임별로 Fast RCNN inference 모델을 적용하여 사용자 인식을 수행하였다. Faster RNN은 Ren et al.(2015)가 제안한 모델로 기존 RNN의 순차적인 계산을 병렬화하고 수평적인 계산을 전개하여 처리 속도를 높여 보다 빠른 객체 탐지가 가능하다. 이후 Faster RCNN 모델을 적용하여 사용자 인식을 수행한 결과, 영상 속 유아(사람)가 사람으로 인식되면 경계 박스를 생성하고 좌표를 반환한다.

본 논문에서는 반환된 좌푯값을 기준으로 Fig. 2(a)와 같이 유아별 데이터를 생성하였다. 밀고 당기기 등의 2명 이상의 유아가 함께하는 행동의 경우에는 각각의 유아 별로 바운딩 박스로 만들어 입력 데이터를 생성하였다.

두 번째 단계로 사용자 인식의 결과를 바탕으로 자세 인식을 수행하여 관절 좌표를 추출하였다. 사용자 인식 결과 사람으로 인식한 위치의 바운딩 박스에 HRNet 모델을 top down 형식으로 적용하였으며 그 결과 사람당 [x, y, keypoint_score] 형식으로 총 17개의 좌표를 추출하였다. HRNet은 Wang et al. (2020)이 제안했다. 이는 고해상도 이미지와 저해상도 이미지의 특성을 동시에 병렬 형태로 추출하여 컨볼루션의 정보 손실을 줄이고 멀티스케일 특성 학습을 통해 고해상도 이미지 처리 성능을 높이려는 목적으로 설계되었다. 이후 생성한 경계 박스를 통해 유아 별로 추출한 프레임에서 관절 좌표를 Fig. 2(b)와 같이 저장할 수 있도록 하였다.

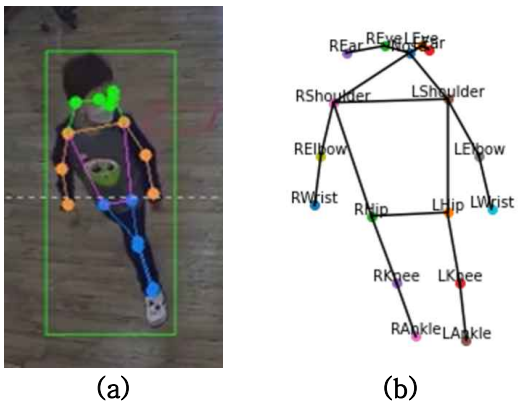

Fig. 2 Input image processing result (a) Human detection and (b) skeleton extraction result

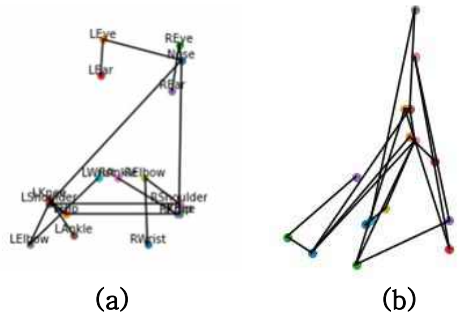

Fig. 3 Pose estimation error examples

세 번째 단계로, 딥러닝의 모델 입력 데이터를 생성하기 위해서 노이즈를 제거하였다. Fig. 3에 나타난 것과 같이 자세 인식의 결과로 얻은 좌표 중 사람의 자세로 볼 수 없을 때 노이즈로 정의 하고 이를 제외하도록 하였다. 입력 영상

에서 전신이 정면으로 보이는 경우에는 입력 영상과 동일하게 포즈 예측에 문제가 없으나, 정면을 보다가 측면을 보거나 뒤로 도는 경우, 서 있다가 앉거나 기는 경우 등에는 좌표 인식의 시작점인 얼굴의 눈, 코 인식이 되지 않고, 어깨 등의 좌표가 인식되지 않아 Fig. 3과 같이 사람의 포즈 예측에 오류가 발생한다. 따라서 관절을 연결한 형태가 사람의 포즈로 인식이 어려운 경우를 노이즈로 판단하여 이를 제거 하였다. 노이즈 판단 기준은 평균 관절 길이 대비 현저하게 짧은 경우로 정의하였으며, 모든 관절 길이를 계산한 후 관절의 길이가 Z-Score 검증(t=1.96) 95%의 신뢰수준에서 벗어난 경우를 노이즈로 간주하여 데이터를 삭제하였다.

이후 유아 별로 입력 데이터를 생성하기 위해서 행동을 정적인 행동과 동적인 행동으로 나누어 행동 종류에 따라 다른 방법으로 데이터를 군집화 하였다. 정해진 위치에서 행동하는 걷기, 서 있기, 비틀기, 밀고 당기기 등의 행동에 대해서는 추출된 모든 관절 좌표에서 코의 좌표를 기준으로 유아가 서 있는 위치별로 K-means clustering 알고리즘을 적용하여 군집화하였다. 코를 기준으로 선정한 이유는 인체의 중심축과 움직임을 파악하는 데 있어 코가 안정적인 기준점이기 때문이다. 유아의 팔다리 움직임은 수시로 변화하기 때문에 유아를 구분하는 기준 점으로 사용하기가 어렵다. 또한 실험에서 유아들이 정면에 움직임을 유도하는 성인을 바라보고 있는 상태로 진행되었기 때문에 코 좌표가 움직임이 가장 적고 고정 되어있기 때문에 데이터를 그룹핑하는 기준으로 사용하였다. 유아 행동이 정해진 위치에서 행해지지 않는 넘기, 달리기, 기어 다니기 등의 동적인 행동의 경우 다중 객체 인식(MOT) 모델을 적용하여 유아별 데이터를 구분하여 군집화하였다(W Luo et al. 2021). 동적인 행동인 넘기, 달리기, 기어 다니기 등의 경우에는 코의 위치가 실시간으로 계속 변화하여 K-means clustering 알고리즘을 통해 군집화한 데이터로 구분이 어렵다. 따라서 유아의 동선을 추적하기 위해 다중 객체 추적 방법을 적용하였다. 이후 유아별로 데이터를 기준으로 하여 행동값을 라벨링하였다.

이후 딥러닝 모델에 입력을 만들기 위하여 유아별로 입력 데이터를 일정 시간 동안 축적하여 프레이밍 하였다. 17개의 추출된 관절 좌표를 coco-dataset 양식에 맞게 matplotlib 모듈을 이용해 연결하여 관절 형태의 이미지를 생성했다(E Bisong et al. 2019). 이미지는 총 영상 속 유아별로 13개의 행동으로 나누어져서 저장되었다. 이후 Fig. 4과 같이 이미지를 30개씩 묶어 프레이밍 하여 (channel x frame x height x width)의 RGB 영상 형태로 gif 파일을 생성하였다. 프레이밍 된 이미지를 딥러닝 모델의 입력하기 전에 데이터 증식을 수행하였다. 총 254개의 트레인 데이터를 프레임별로 같은 위상각 변이를 적용해 gif 파일마다 -30도에서 30도까지 랜덤하게 이미지 각도를 조정하여 저장하였다. 최종적으로 그룹화하여 획득한 데이터는 Fig. 4와 같다.

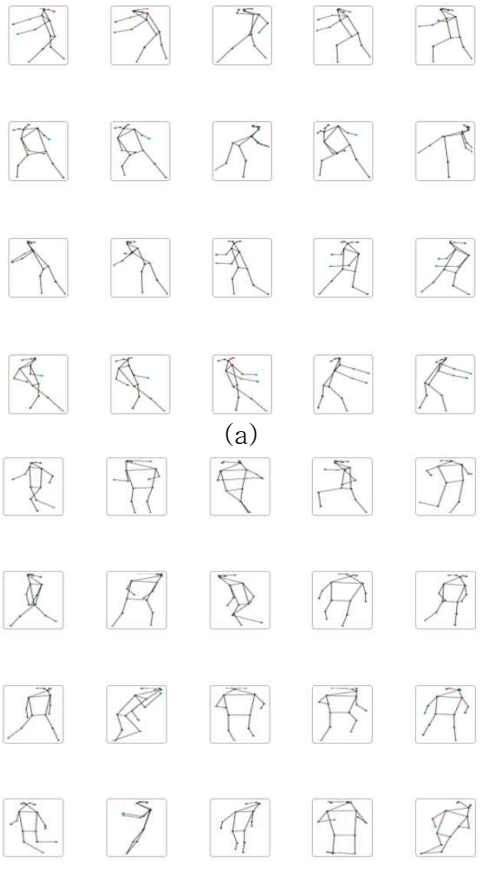

Fig. 4 Input data sequence (a) Push and pull sequence (b) Running sequence

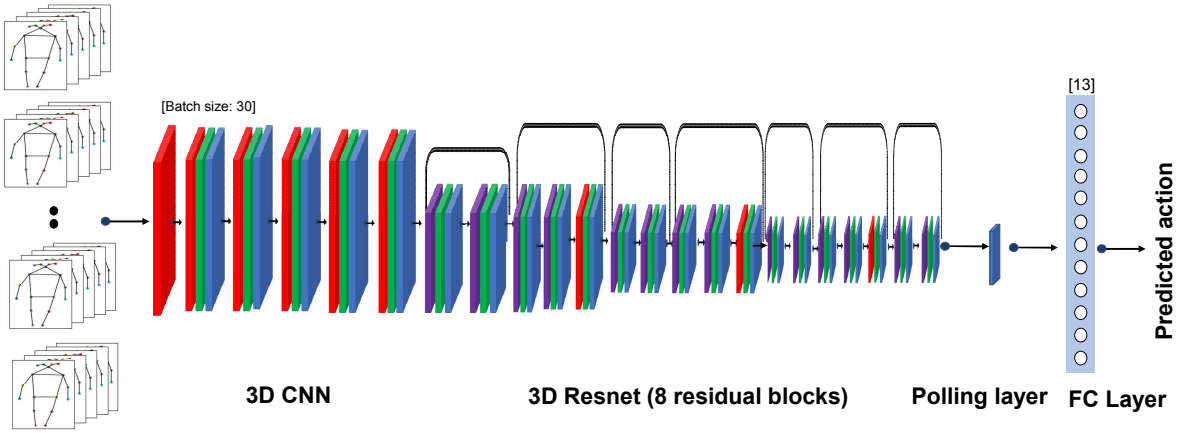

Fig. 5 3D ResNet architecture

2.2 제안하는 딥러닝 모델

본 연구에서 개발한 딥러닝 모델은 Fig. 5와 같이 구성되었다. 사용한 행동 인식 모델은 Hara, K et al.(2017)가 제안한 3D Resnet 모델을 사용하였다. 3D ResNet은 잔차를 학습하는 2DResNet을 기반으로 하여 공간정보와 시간 정보를 모두 사용하는 3D 합성공 신경망을 이용한다. 많은 연구에서 3DResnet 모델을 사용한 이유는 영상의 공간 및 시간 정보를 동시에 고려할 수 있으며, 깊이가 매우 큰 네트워크를 효율적으로 학습하는 잔차 학습을 통해 복잡한 패턴을 학습하는데 유용하기 때문이다. 본 연구에서는 사전훈련된 3D Resnet 모델을 사용하였고 전이 학습을 통해 현재 생성한 데이터에 맞는 모델로 미세 파라미터를 조정하였다. 본 실험에서 인식할 13가지의 행동을 저장한 관절 데이터를 3D ResNet 모델의 입력으로 유아 행동을 인식하였다. 3D Resnet은 Kinectics 데이터로 사전 훈련된 모델로 클래스 수는 200개이므로 본 연구의 클래스 수인 13개와 다르다. 때문에 한 데이터 셋에서 학습된 모델의 지식을 다른 데이터 셋에 적용하는 전이학습 기법을 활용하여 마지막 출력 레이어를 13개의 클래스 수에 맞게 수정하였다. 또한 사전 훈련된 기존 모델의 입력 데이터와 스켈레톤 이미지의 형식을 동일하게 맞춰주기 위하여 스켈레톤 이미지 데이터를 RGB 형식으로 변환하였고 재 학습을 통해 모델의 파라미터를 최적화 하였다.

제안한 모델은 64, 128, 256, 512의 채널과 8단계의 잔차 블록으로 구성된다. 각 채널 단계에는 배치 정규화 층과 ReLU 층이 있으며, 이후 컨벌루션 층을 형성한다. 입력 데이터는 (배치 크기, 채널, 프레임, 높이, 너비)의 4D 텐서 형태이다. 초기 컨볼루션의 첫 번째 레이어(Conv3d)는 3개 입력 채널과 64개의 출력 채널, 3x3x3 커널 크기, 1x1x1의 스트라이드 값을 가지며 그 뒤에 배치 정규화 및 ReLU 활성화가 이어진다. 이후 64개의 입력 채널과 64개의 출력 채널 및 배치 정규화와 ReLU 정규화가 진행된다. 이후 아키텍처의 핵심인 잔차 블록이 있는 ResNet 레이어를 통해 신호가 전달된다. 각 레이어에는 2개 이상의 컨벌루션 층이 있으며 층마다 배치 정규화와 Relu 활성화 층이 이어진다. 두 번째 잔차 블록 이후로 128개의 채널 출력하는 Conv3D 컨볼루션층과 배치 정규화, ReLU활성화 층이 두 번 연속으로 이어지며 해당 컨볼루션의 출력 크기를 절반으로 줄인다. Conv3DNoTemporal은 시간적 차원(프레임)을 유지하고 출력 크기, 즉 너비와 높이만 줄여 시간적 정보를 보존하고 공간적 정보만 축소하여 데이터에서 중요한 시간적 패턴을 유지할 수 있도록 한다. 또한, 계산 복잡도를 감소시켜 빠른 연산을 가능하게 하며 모델 파라미터 수를 감소시켜 과적합을 방지할 수 있다. 이후 2개의 잔차 블록을 지날 때마다 컨볼루션의 출력 채널

수는 2배로 증가하고 컨볼루션의 폭과 길이는 2배로 감소한다. 마지막 8번째 잔차 블록을 거치면 컨벌루션은 (30, 512, 30, 4, 4)의 형태를 보이게 되며 Adaptive Average Pooling을 통해 (30, 512, 1, 1, 1) 형태의 풀링 결과를 갖게 된다. 마지막으로 Fully connected layer를 출력 클래스의 수 만큼 설정하여 분류를 수행한다.

3. 실험결과

3.1 데이터 수집 방법

본 연구에서는 2022년 2월 23일에 50명의 유아로부터 13개의 행동을 연속적으로 2회씩 수행하도록 하였으며 3개의 카메라를 통해 이를 녹화하였다 (IRB no. 1044396-202007-HR-138-02). 입력 영상으로부터 158,070개의 이미지 데이터를 수집 하였다. 행동은 Straudenmayer et al. (2009) 의 연구에서 제안된 유아들이 자주 하는 행동 종류에 따라 수집, 분석하였으며 서 있기, 걷기, 구부리기, 비틀기, 넘기, 통에서 물건 꺼내기, 밀고 당기기, 달리기, 제자리 점프하기, 발차기, 기어 다니기, 엎드려있기, 앉아 있기의 총 13개의 행동을 인식하였다. 행동 중 밀고당기기는 2명의 유아가 한 조를 이루어 진행되었으며 데이터는 유아 개인별로 만들어진 경계 상자로 나누어 저장하였다. 자료수집 환경 및 예시는 Fig. 6 에 나타난 것과 같다.

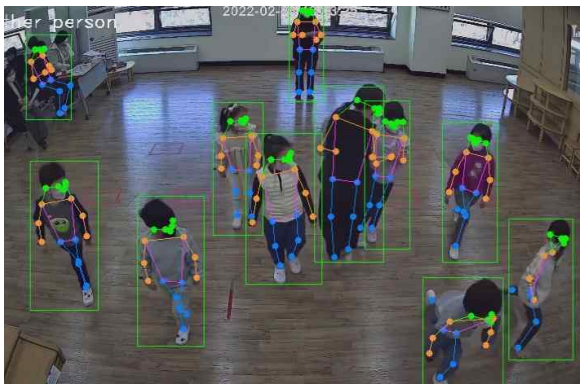

Fig. 6 Data collection environment and examples

3.2 결과분석

결과 분석을 위하여 전체 데이터를 추출한 후 행동 별로 60%의 데이터를 모델 학습에 사용하였고, 20%를 모델 확인에 20%의 데이터를 테스트에 사용하였다. 전체 실험 영상에서 유아들은 정해진 13개의 행동을 10초가량 순서대로(일어서 있기 ~ 앉아있기) 수행하였고, 유아 별로 2회씩 전체 동작을 반복하였다. 전체 획득된 데이터의 프레임 수는 총 5,758개 이며, 각 행동별로 약 400개 정도의 프레이밍한 데이터가 생성이 되었다. 본 연구에서 제안한 3DResnet 알고리즘을 활용한 유아 행동 인식 알고리즘의 정확도는 평균 72.21%이다. 동일한 데이터를 2D ResNet50 모델에 적용하였을 때 전체 평균 정확도는 55.0%로 획득되었다. 따라서 제안한 방법이 유아들의 행동 인식에 좀더 나은 성능을 보임을 알 수 있었다.

Table 1 Action list and accuracy

Action	Label	# of data	Accuracy
stand	0	14,700	90.74%
walk	1	15,420	61.11%
bend/bow	2	12,900	68.52%
twist	3	13,350	66.67%
go over	4	12,360	64.67%
pick	5	13,380	64.81%
push/pull	6	17,940	88.89%
run/jog	7	12,180	62.96%
jump/leap	8	12,480	77.78%
kick	9	14,220	33.33%
crawl	10	10,890	75.93%
lay down	11	8,250	92.59%
sit	12	14,670	90.74%
Proposed method			72.21%
2D ResNet50			55.0%

13개의 행동 중 서 있기(90.74%), 걷기(68.52%), 구부리기(85%), 비틀기(66.67%), 넘기(64.67%), 통에서 꺼내기(64.81%), 밀고 당기기(88.89%), 달리기(62.96%), 제자리 점프하기(77.78%), 발차기(33.33%), 기어 다니기(75.93%),

엎드려있기(92.59%), 앉기(90.74%)의 정확도를 보였다. 서있기나 앉기, 엎드려있기의 경우 정확도가 90% 이상으로 높았으며, 밀고 당기기, 제자리 점프하기, 기어 다니기 순으로 정확도가 높게 획득되었다. 일어서 있는 경우는 데이터 수집 전후에 가장 많은 데이터가 획득되어 일어선 경우의 데이터가 잘 학습이 되었음을 확인할 수 있었으며 앉거나 눕는 자세도 안정적으로 높은 정확도로 인식이 되었다. 밀고 당기기나 뛰기, 기기 등과 같이 행동이 다른 행동과 뚜렷하게 구분이 될 때도 88.89% ~ 77.78%의 정확도를 보였다. Fig. 7의 confusion matrix 분석에 따르면 발을 차는 행동에 걷기, 뛰기 등의 행동이 중간에 들어가 있어 프레임의 입력만으로 분석이 어려운 문제가 있었다. 발차기의 경우 넘기나 서기 등과 유사한 행동 이미지가 저장되어 타 행동과 구분이 안 되는 문제가 있었으며 때문에 낮은 33.33%의 정확도를 보였다. 통에서 물건을 꺼내는 행동의 경우에도 구부리기나 비틀기, 뛰기 등으로 잘 못 예측이 되는 것을 확인할 수 있었는데 통에서 물건을 꺼내는 행위가 몸을 비트는 동작이 중간에 들어가게 되면 이를 잘못 인식하는 것으로 볼 수 있다.

Fig. 7 Confusion matrix result

본 연구의 제약점으로는 유아 행동 데이터가 시간에 따라 변화 하면서 다른 종류의 행동들과 구분이 되지 않는 특징이 있는 경우 정확도가 낮은 한계가 있었다. 특히, 발차기의 경우 서있다가 발차는 행동이 이루어지므로 걷기와 유사하게 판단이 되거나 중간에 서있는 시간이 길어지면 발로 차는 짧은 시퀀스를 인지하지 못하는 경우가 발생하였다. 또한, 데이터 수집 시 측정 카메라의 위치가 정면이 아닌 측면 2개 카메라의 경우에는 얼굴 정보가 획득되지 않아 포즈 인식에 한계가 있었다. 따라서 향후에는 데이터 증강 기법을 통해 정면이 아닌 측면 영상 데이터를 추가로 확보하거나 데이터 수집 단계에서 얼굴이 나오는 측면의 데이터를 확보하는 것이 필요하다. 본 연구에서 목표로 하는 자유로운 환경에서 측정한 데이터 분석으로 인해 움직임에 대한 노이즈나 입력 신호의 시작 시점과 종료 시점이 행동에 따라 다른 부분에 대해서도 향후 추가적 연구가 필요하다. 또한, 영상만으로 행동을 인식하므로 걷기나 달리기, 뛰기와 서있기 등의 행동은 영상으로는 구분이 어렵고 가속도나 각속도 등의 속도 변화를 추가로 감시하여 분석하는 것이 필요하다. 마지막으로 유아들의 경우에는 동일한 지시 행동의 경우에도 행동이 부정확하거나 불규칙한 예도 있었으며 이러한 부정확한 행동을 안내하는 성인과 유아의 접촉이 잦아 관절 좌표 추출 시 제외하는 것에도 한계가 존재했다. 따라서 이를 보완할 수 있는 데이터에 대한 전처리가 필요하며, 다양한 패턴의 행동에 대한 학습이 이루어질 수 있도록 다양한 종류의 입력에 강건한 딥러닝 모델에 대한 개발이 필요하다 하겠다.

4. 결 론

본 연구에서는 다수의 유아가 등장하는 영상 내의 행동을 인식하기 위하여 딥러닝 기반의 유아 행동 인식 기술을 개발하였다. 유아들의 경우 동일한 행동이라도 표현과 방법이 다양하여 다양한 종류의 입력에 강건하게 분석될 수 있는 딥러닝 모델에 대한 개발이 필요하다. 본 연구에서는 입력 신호를 딥러닝의 입력에 맞도록 처리하고 3D ResNet을 사용하여 행동 인식 알고리즘을 제안하였다. 실험결과 13개의 행동 인식에 평균 72.21% 정확도를 보였으며, 행동 중 서

있기, 밀고 당기기, 앉기 등의 행동 경우 90% 이상의 높은 인식률을 보였다. 그러나 발차기 등의 행동은 낮은 인식률을 보였으며, 향후 센서 데이터 등과의 통합을 통해 속도가 중요한 행동의 경우 속도 정보를 개선할 수 있도록 하고자 한다. 또한, GCN, X3D, PoseC3D등 여러에 사용되는 SOTA모델을 적용하여 다양한 패턴의 유아 행동을 인식하기 위한 모델을 통해 정확도를 개선하고 GAN 등의 데이터 생성 기술을 활용하여 데이터가 충분하지 않거나 부족한 경우 활용이 가능하도록 하고자 한다.

References

Bisong, E., and Bisong, E. (2019). Matplotlib and seaborn. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, 151-165.

Chen, K., Wang et al. (2019). MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155.

Duan, H., Zhao, Y., Chen, K., Lin, D., and Dai, B. (2022). Revisiting skeleton-based action recognition. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* June 18-24, LA, USA, pp. 2969-2978.

Hara, K., Kataoka, H., and Satoh, Y. (2017). Learning spatio-temporal features with 3d residual networks for action recognition. *In Proceedings of the IEEE international conference on computer vision workshops*, Oct. 22-29, Venice, Italy, pp. 3154-3160.

Jobanputra, C., Bavishi, J., and Doshi, N. (2019). Human activity recognition: A survey. *Procedia Computer Science*, 155, 698-703.

Jung, D. J. and Yun, J. O. (2011) Human Activity Recognition using Model-based Gaze Direction Estimation, *Journal of the Korea Society Industrial Information System,* 16(4), 9-18.

Kale, G. V. (2019) Human activity recognition on real time and offline dataset. *Int. J. Intell. Syst. Appl. Eng*. 7(1), 60, 8211-65.

Kim, K., Jalal, A., and Mahmood, M. (2019). Vision-based human activity recognition system using depth silhouettes: A smart home system for monitoring the residents. *Journal of Electrical Engineering & Technology*, 14, 2567-2573.

MMCV Contributors. 2018. MMCV: OpenMMLab Computer Vision Foundation. https://mmcv.readthedocs.io/en/2.x/get_started/introduction.html.

Olalere, F., Brouwers, V., Doyran, M., Poppe, R., and Salah, A. A. (2021). Video-Based Sports Activity Recognition for Children. *In IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, December 14-17, Tokyo, Japan, pp. 1563-1570.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* June 15-20, CA, US, pp. 12026-12035.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Shahroudy, A., Liu, J., Ng, T. T., and Wang, G. (2016). Ntu rgb+d: A large scale dataset for 3d human activity analysis. I*n*

Proceedings of the IEEE conference on computer vision and pattern recognition, June 27-30, Las Vegas, NV, USA, pp. 1010-1019.

Staudenmayer, J., Pober, D., Crouter, S., Bassett, D., and Freedson, P. (2009). An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of applied physiology*. 107(4), 1300-1307.

Wang, J. et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10), 3349-3364.

Zhang, H. B., Zhang, Y. X., Zhong, B., Lei, Q., Yang, L., Du, J. X., and Chen, D. S. (2019). A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5), 1005.

박 재 석 (Jaeseok Park)

- 가천대학교 산업경영공학과 학사과정
- 관심분야: 영상처리, 모바일 헬스케어, 머신러닝, 딥러닝

차 기 주 (Kijoo Cha)

- 이화여자대학교 유아교육학과 문학사
- 하버드대학교 교육대학원 (Harvard Graduate School of Education) 국제교육정책 석사(Ed. M.)
- 스탠포드대학교 교육대학원(Stanford Graduate School of Education) 철학박사(Ph.D.)
- (현재) 가천대학교 사회과학대학 유아교육학과 부교수
- 관심분야: 유아교육 공간과 놀이행동, 유아교육정책, 사회정서발달 등

최 아 영 (Ahyoung Choi)

- 정회원
- 이화여자대학교 정보통신학과 공학사
- 광주과학기술원 정보통신학과 공학석사
- 광주과학기술원 정보통신학과 공학박사
- 삼성전자 무선사업무 책임 연구원
- (현재) 가천대학교 IT대학 AI.소프트웨어학부 부교수
- 관심분야: 생체신호처리, 모바일 헬스케어, 머신러닝, 딥러닝