2021년 2월

박사학위논문

앙상블 및 관심영역기반 심층신경망을 이용한 행동인식과 행동특성분석

조 선 대 학 교 대 학 원

제 어 계 측 공 학 과

변 영 현

앙상블 및 관심영역기반 심층신경망을 이용한 행동인식과 행동특성분석

Action recognition using ensemble-based and ROI-based deep neural network and analysis of action characteristics

2021년 2월 25일

조 선 대 학 교 대 학 원

제 어 계 측 공 학 과

변 영 현

앙상블 및 관심영역기반 심층신경망을 이용한 행동인식과 행동특성분석

지도교수 곽 근 창

이 논문을 박사학위신청 논문으로 제출함.

2020년 10월

조 선 대 학 교 대 학 원

제어계측공학과

변 영 현

변영현의 공학박사학위논문을 인준함

위원장 조 선 대 학 교 교수 반 성 범 인

위 원 조 선 대 학 교 교수 신 주 현 인

위 원 조 선 대 학 교 교수 염 홍 기 인

위 원 한국전자통신연구원 박사 김 도 형 인

위 원 조 선 대 학 교 교수 곽 근 창 인

2020년 12월

조선대학교 대학원

목 차

표 목차

그림 목차

ABSTRACT

Action recognition using ensemble-based and ROI-based deep neural network and analysis of action characteristics

Byeon, Yeong Hyeon
Advisor : Prof. Kwak, Keun Chang, Ph. D.
Dept. of Control and Instrumentation Eng.,
Graduate School of Chosun University

This paper conducts behavior recognition and behavioral characteristic analysis using deep neural networks based on ensembles and regions of interest. Video-based behavior recognition is a method of automatically identifying the behavior displayed by a target person through digital data processing. It can be applied to video-based automatic crime monitoring, automatic sports video analysis, and whether the situation of a silver robot. In particular, as the necessity for silver robots increases to better care for the elderly due to the aging of society, research on behavior recognition as a core technology is also becoming more important. Behavior recognition data is mainly composed of images and skeletons, and recognition performance can be improved by combining the analysis of data with different characteristics. In addition, the image data used for behavior recognition is composed of sequences as well as spatial information and contains time

information. Therefore, performance can likely be improved by analyzing and combining spatial or temporal information in an optimal structure. Important information can be gleaned from behavior recognition data if the region of interest is correctly placed on the person performing the action, thus removing ambient noise, and if the behavior recognition algorithm conducts feature analysis by focusing on the behavior itself rather than learning the entire domain. In addition, since humans use tools to perform actions differently from animals, training of a neural network by placing a region of interest on a hand-object enables feature analysis by focusing on tool-related information. Performance can be improved by combining information from models that have been trained by focusing on these regions of interest. Because physical conditions change with age, the characteristics of the data differ according to the age of the actor. To analyze the behavioral characteristics of these differences, an explainable artificial intelligence technique can be used. The database used for this experiment is the ETRI-Activity3D database, which contains color images, skeletons, and depth images of 55 daily behaviors of 50 elderly individuals and 50 adults. In this experiment, the performance of the proposed models, the RGB-S-based 3-stream ensemble model and the ROI-based ensemble model, improved by at least 2.6% and up to 20.97% compared to other behavior recognition methods. In addition, the heat trajectory was obtained from the skeleton information through an explainable artificial intelligence technique, and comparative analysis was performed with the RGB video.

제1장 서론

제1절 연구 배경

현대사회는 나이를 먹으면서 손상되는 신체기능들을 의학, 공학 등의 기술을 통해 일정수준까지 복구시켜 건강을 보존시킬 수 있게 되었다. 이는 과거시대보다 현대인들의 수명이 길어지게 되어 노인인구의 증가를 의미한다. 또한 출산율 저하로 인하여 새로 태어나는 인구가 적고 기존 청년들이 노인세대로 계속 유입되면서 고령인구는 급격하게 증가하고 있다. 인구의 고령화에 따라 고령인구수 대비 청년인구수가 현저히 줄어들 경우 사회 전반에 혼란이 예상된다. 이러한 노인의 문제가 현대사회에 와서 더 문제가 되는 데에는 사회의 구조적 변화의 측면도 있다[1-3].

과거사회는 생계를 유지하는데 중요한 의식주 해결에 급급하여 사람들 대부분이 농업, 상업, 어업 등 노동이 중요한 직업군을 주로 가졌고 그러한 경제활동은 사회의 구성단위가 대가족이 되도록 하였다. 대가족 체제는 나이가 들어서 몸이 불편해지고 경제활동이 불가능해진 노인의 보살핌을 대체로 가족 구성원에게 분담시키는 특징이 있다. 하지만 현대 사회는 과거대비 의식주가 풍족해지고 직업군이 다양해져서 더 이상 대가족 체제의 의존성이 줄어들게 되었고, 핵가족화가 진행되면서 가족 구성원이 흩어짐으로 인해 노인 보살핌은 사회적 역할로 돌려지고 있다[4,5]. 노인 보살핌은 반복적 노동이고 상황에 따라 힘들기 때문에 사회적으로 인간이 아닌 가정용 실버 로봇이 대신 일을 할 수 있도록 연구 개발에 힘쓰고 있다. 가정용 실버 로봇의 경우 공장 제조 로봇의 단순 움직임과 달리 마주하는 환경이 복잡해서 적절한 대응을 위해서는 고도의 인공지능 기술을 필요로 한다[6-10].

행동인식 기술은 카메라, 관성센서 등의 입력 데이터를 분석하여 행위자가 어떤 행동을 하고 있는지 자동으로 알아내는 기술이다. 행동인식을 통해 어떤 사람의 행동을 알아냄으로서 그 사람이 처한 환경을 이해하고 그에 알맞은 반응을 할 수 있다. 예를 들면, 가정용 실버로봇이 홀로 있는 노인을 항상 주시하면서 갑작스런 쓰러짐과 위험들을 판단하고 그에 적절한 이벤트들을 수행할 수 있어 노인 보살핌의 사회문제가 수월해진다[11,12].

노인은 젊을 때 보다 뼈의 밀도가 감소하게 되어 심해지면 뼈가 쉽게 부러질

수 있다. 여자의 경우 나이를 먹으면서 폐경이 진행되어 뼈의 분해를 지연시켜주는 에스트로겐 양이 줄어들고 이는 더 빠른 골밀도 감소로 이어진다. 또한, 비타민 D가 감소하고 칼슘 흡수율이 줄어들어 척추뼈, 손목의 팔뼈, 고관절의 허벅지뼈가 약해진다. 척추 위쪽 뼈가 약해지면 머리가 앞으로 기울어 입안이 눌러지고 삼키는 것이 어려워져 질식의 위험이 증가한다. 늙어가면서 척추뼈의 밀도가 감소하고 척추뼈 사이의 디스크가 닳아 없어져 척추의 길이가 줄어든다. 이로 인해 노인들은 키가 작아진다. 관절 중간의 연골은 오랜 시간의 움직임으로 닳아서 얇아지고 관절이 잘 미끄러지지 않게 되어 관절의 손상이 빨라진다. 이는 노년에 골관절염을 흔하게 일으킨다[13,14]. 늙어가면서 사람들의 인대와 힘줄은 탄력이 떨어져 유연성이 떨어지고 쉽게 손상되며 회복도 느려진다. 30대를 넘어서면서 신체활동이 줄어들고 테스토스테론과 성장호르몬이 줄어들어 근육량과 근력이 계속 감소한다. 또한, 빠른 수축의 근섬유의 소실이 느린 수축의 근섬유보다 더 줄어들어 움직임 속도가 느려진다. 노화로 인해 약 10%에서 15% 이하의 근육이 손실되고 이는 운동을 통해 예방 할 수 있다. 대부분 노인들은 일상 움직임에 충분한 근육을 유지하지만 어느 정도 저하를 느끼고 체지방의 비율도 변해서 체형이 바뀐다[15]. 노화로 인해 수정체가 굳어져 근처의 물건에 초점을 맞추는 것이 힘들어지고, 수정체 밀도가 증가하여 어두운 곳에서 잘 보지 못하며 동공이 빛에 늦게 반응한다. 그리고 수정체가 뿌옇게 변해서 색상 식별이 어렵고 신경 세포 수가 감소하여 거리 감각이 줄어든다[16]. 큰 소음에 오래 노출되면 귀의 청력이 손상되기도 하지만 노화함에 따라서도 청력이 손실된다. 노인성 난청으로 높은음의 소리를 잘 못들어서 단어의 이해가 어려워지는데 그 이유는 자음이 높은음인 경우가 많고 단어를 구분하는 데 중요한 소리이기 때문이다[17]. 나이를 먹으면서 뇌의 화학 메시지의 수용체 중 일부가 손실되고 뇌로 가는 혈류가 감소하여 일부 기능을 잘하지 못하게 된다. 그래서 노인들은 반응을 느리게 하지만 시간이 충분하면 주어진 일을 잘한다. 그리고 척수 세포수가 줄어들어 노인의 감각이 둔화된다[18]. 노화에 따라 호흡에 사용되는 근육이 약해지고 폐의 모세혈관 수가 감소하여 산소 흡수율이 낮아져 힘든 움직임을 어렵게 한다. 소화계는 노화에 영향이 적은 편으로 위장에서 음식이 느리게 소화되고, 위장이 약해서 음식물을 많이 담고 있지 못한다[19]. 사람은 노화함에 따라 신체적 능력이 변하여 젊을 때와 다른 행동특성을 보인다. 성인과 노인의 행동특성분석을 하여 실버로봇이 정확한 판단에 도움이 될 수 있고 노인에게 맞춤형 서비스가 가능하다.

제2절 연구 목적

행동인식 데이터는 신호의 종류가 다양하고 크기도 커서 복잡하기 때문에 단일 모델만으로 풍부한 표현의 특징을 충분히 추출하지 못하는 문제가 있다. 그러므로 다방면의 데이터 특성에 입각하여 적절한 모델을 세분화하고 특징을 다양하게 분석할 필요성이 있다. 3D-CNN의 시공간 특징뿐만 아니라 추가적인 모델들의 설계를 통해 시공간 특징을 다양화시켜 최종적으로 앙상블하는 모델을 제안한다. 추가적인 모델은 RGB이미지 또는 스켈레톤을 입력으로 시간정보에 더 집중하거나 공간정보에 더 집중하는 구조의 모델로 구성하였다. 둘째로, 행동인식 데이터는 3차원으로 데이터 크기가 커서 불필요한 정보도 배로 증가하는 문제가 있었다. 이 문제를 해결하고 인식률을 높이기 위해 행동에 중요한 사람부분과 행동의 주요 정보인 손의 도구(hand-object)에 집중하는 모델을 설계하여 특징을 다양화시키고 최종적으로 앙상블 하는 모델을 제안하였다. 셋째로, 사회의 고령화 속에서 노인과 성인의 신체적 차이에 따라 행동특성도 달라지는데 인지과학적으로 행동특성분석을 하는 연구가 적은 문제가 있다. 이를 해결하기 위해 설명가능한 AI(Artificial Intelligence)를 이용해 행동특성분석 방법을 제안하였다.

기존 행동인식의 동향은 다양한 입력데이터로 여러 심층 신경망 모델을 설계하고 결합하여 성능을 개선시키고 있다. 행동인식 데이터는 주로 이미지와 스켈레톤으로 구성되어 특성이 서로 다른 데이터의 분석을 결합하면 더 좋은 인식 성능을 기대할 수 있다. 또한 행동인식의 이미지 데이터는 공간 정보뿐만 아니라 시퀀스로 구성되어 시간정보도 가지고 있다. 그래서 공간정보 또는 시간정보에 각각 최적의 구조로 분석하고 결합함으로서 더 좋은 성능을 기대할 수 있다.

행동인식에서 중요한 정보는 행동을 수행하는 사람 그 자체로 주변 잡음을 제거하고 사람에 관심영역을 두어 신경망을 학습시키면 전체 영역을 학습할 때보다 행동자체에 집중하여 특징분석이 가능하다. 또한 사람은 동물과 다르게 행동을 수행하는데 도구를 사용하므로 손의 물체에 관심영역을 두어 신경망을 학습시키면 도구정보에 집중하여 특징분석이 가능하다. 이러한 관심영역들에 대해 집중하여 학습된 모델들의 정보를 결합하면 더 좋은 성능을 기대할 수 있다.

행동인식에서 행위자의 연령에 따라 데이터의 특성에 차이를 보인다. 크게 성인과 노인의 행동을 분석해보면 성인은 노인보다 신체적 균형을 잘 잡고 외부 작

용에 신속하게 반응할 수 있고 모션이 빠른 반면 노인은 관절염, 허리 디스크, 근육량 저조 때문에 외부 작용에 반응이 더디고 모션이 느리다. 그리고 노인은 신체 균형 능력이 떨어져 낙상공포 심리를 갖고 시각의 기능이 많이 떨어져 얼굴을 물체에 가까이 대거나 보조 장비를 사용해야 한다. 노인은 학습능력이 뒤떨어져 최신 기기 사용에 시행착오를 많이 겪고 치아가 썩고 닳아서 저작을 천천히 오래해야 한다. 이러한 차이에서의 행동특성을 분석하기 위해 설명 가능한 인공지능 기법을 이용할 수 있다. 따라서 본 논문은 3가지 방향으로서 비디오와 스켈레톤 기반 심층 신경망 모델들을 결합하고 관심영역기반 행동인식 모델을 설계하고 설명 가능한 인공지능을 이용한 행동특성분석을 제안한다.

제3절 연구 내용

실험을 위해 사용한 데이터셋은 ETRI-Activity3D로 50명의 노인과 50명의 성인의 일상적인 55개 행동에 대한 컬러이미지, 스켈레톤, 뎁스이미지들을 포함하고 있다. 이 데이터셋은 총 112,620개 샘플로 구성되어 있는 두 번째로 큰 규모의 행동인식 데이터로 실제 주거환경에서 최대 8대의 다방면의 키넥트 v2를 이용해 취득하였다. 또한 실버로봇 환경을 가정하여 센서는 70cm와 120cm 높이로 구성하여 1.5~3.5m 거리 내에서 취득하였다.

첫 번째 제안의 실험결과로서 RGB비디오 입력의 단일 모델인 2D-CNN-LSTM-Type1, 2D-CNN-LSTM-Type2, 3D-CNN의 정확도는 각각 49.47%, 47.49%, 79.20%가 나왔고 스켈레톤 입력의 단일 모델인 PEI(Pose Evolution Image)-T1-2D-CNN, PEI-T2-2D-CNN, PEI-T3-2D-CNN, PEI-T4-2D-CNN의 정확도는 각각 84.95%, 85.88%, 86.09%, 85.20%가 나왔다. 앞서의 단일 모델 중에 RGB비디오 입력의 단일 모델인 2D-CNN-LSTM-Type1, 3D-CNN과 스켈레톤 입력의 단일 모델인 PEI-T3-2D-CNN이 모두 앙상블 된 제안한 RGB-S-based ensemble network (Type2)의 정확도는 93.20%로 단일 모델대비 최소 7.11%에서 최대 43.73%의 정확도가 개선되었다.

두 번째 제안의 실험결과로서 Body ROI 입력의 3D-CNN과 Hand-object ROI 입력의 3D-CNN의 정확도는 각각 76.85%, 73.11%를 보여주었다. RGB비디오 입력의 단일 모델인 3D-CNN, BodyROI-3D-CNN, HandObject-3D-CNN과 스켈레톤 입력의 단일 모델인 PEI-T3-2D-CNN이 모두 앙상블 된 제안한 ROI-based ensemble network (Type6)의 정확도는 94.87%로 단일 모델대비 최소 8.78%에서 최대 21.76%의 정확도가 개선되었다. 또한 다른 연구의 방법들과 비교했을 때, 제안한 ROI-based ensemble network (Type6)이 최소 4.27%에서 최대 20.97%의 정확도가 개선되었다.

세 번째 제안의 실험결과로서 제안한 방법을 통해 히트케적을 구하고 그 결과를 RGB비디오와 비교분석하였다. 음식을 먹는 행동의 단일 데이터 비교에서 한손으로 음식을 먹고 다른 손은 무릎에 얹어놓는 행동특성이 히트케적에서 확인되었고 또한 음식이 입 근처에 왔을 때 목을 살짝 당겨 먹는 행동특성이 히트케적에서 확인 되었다. 그리고 음식을 먹을 때 한 손으로는 음식을 들고 나머지 손으로 음식을 흘리지 않기 위해 받치는 행동특성이 히트케적에서 확인되었다. 음식을 먹는

행동의 노인과 성인의 데이터 비교에서 노인은 양손과 목을 주로 활용하는 반면 성인은 양손과 목, 어깨, 팔꿈치까지 상반신을 전체적으로 활용하는 것이 히트케적에서 확인되었다.

본 연구를 통해 기여한 부분은 첫째로 다양한 입력으로 시간 또는 공간에 집중된 모델들을 통해 특징을 다양화시키고 앙상블하여 행동인식 정확도를 높였고, 둘째로 사람 또는 손의 물체에 관심영역을 적용하여 불필요한 정보의 제거를 통해 행동인식의 핵심정보에 더 집중시킴으로써 특징을 다양화시키고 최종적으로 앙상블하여 정확도를 높였다. 셋째로 설명가능한 AI 방법을 통해 노인과 성인의 행동 데이터로부터 행동특성을 분석하는 방법을 제시하였다.

본 논문은 앙상블 및 관심영역기반 심층 신경망을 이용한 행동인식과 행동특성 분석을 연구한다. 2장은 노인의 행동특성분석을 나타내고 3장은 딥러닝을 이용한 행동인식을 기술한다. 4장은 제안하는 행동인식 방법을 나타내고 5장은 실험 및 결과분석을 기술한다. 끝으로 6장은 결론을 맺는다.

제2장 관련 연구 동향

제1절 행동인식 동향조사

스켈레톤 기반 행동인식에 대한 연구가 진행되어 왔다. 최근 RNN은 Gradient vanishing, exploding 문제와 장기기억을 학습하는데 어려움이 있다. 이를 해결하기 위해 LSTM과 GRU(Gated Recurrent Unit)이 개발되었지만 그것의 하이퍼 탄젠트와 시그모이드 함수의 사용은 레이어들에 대한 Gradient decay 문제를 낳았다. 그래서 Li[20]는 같은 레이어의 뉴런들은 독립적이면서 레이어간에는 연결된 indRNN을 제안하였다. 이 네트워크는 기존 RNN보다 더 깊게 쌓을 수 있고 더 긴 시퀀스를 처리할 수 있게 되었다. 이 네트워크를 NTU RGB+D데이터셋을 이용하여 스켈레톤 기반 행동인식을 수행하였다. 스켈레톤 기반 행동인식의 경우 제한된 특징 표현력으로 큰 데이터셋에서 한계가 있고 근래의 RNN은 기하 관계를 무시하고 시간에 따른 신체 관절의 변화에 초점을 맞추어 접근하고 있다. Wang[21]은 행동인식을 위해 관절간의 기하관계를 반영하기 위해 관절, 엣지(edge), 표면(surface)를 소개하였다. 이 3개의 기하정보를 일반적인 RNN에 입력으로 사용하며 시점 변환 레이어(viewpoint transformation layer)와 시간적 드롭아웃 레이어(temporal dropout layer)를 사용하였다. 그리고 행동검출을 위해 프레임별 행동을 분류하여 멀티스케일 슬라이딩 윈도우 알고리즘을 이용하였다. 최근 스켈레톤을 이용한 행동인식 방법들은 대부분 RNN기반이다. Li[22]는 행동인식과 행동검출을 위한 새로운 CNN을 제안하였다. 로우 스켈레톤 좌표와 스켈레톤 모션이 CNN에 입력으로 들어간다. 중요한 스켈레톤 관절을 자동으로 선택하고 정렬하기 위한 새로운 스켈레톤 변환 모듈이 설계되었다. 행동검출을 위해 시간적 부분 제안을 추출하는 윈도우 제안 네트워크를 개발했다. 사람 신체 스켈레톤의 역학은 행동인식을 위한 중요 정보를 가진다. 스켈레톤을 모델링 하기 위한 기존 방법들은 손수 제작된 부분에 의존하였다. 그러므로 제한된 표현능력과 일반화의 어려움이 따랐다. 그래서 Yan[23]은 이미지에서 프레임별 스켈레톤 정보를 추출하고 그 정보를 시간적 차원을 가지는 스켈레톤 그래프로 나타내어 ST-GCN(Spatial Temporal Graph Convolutional Network)으로 분류하였다. 관절의 계층구조와 의미론적 규칙은 행동인식을 위해 중요한 정보를 포함한다. 기존 GCNN 방법들은 스켈레톤 구조를 모델링 하기 위해 물리적으로 연결된 이웃점들만 고려하였다. 그래서 고차원적 정보를 추출하는데 실패한다. Wen[24]은 스켈레톤 시퀀스의 다른 범위에 대한 지역적 시간정보를 이용하여 계층적 공간 구조의 특징을 추출하고 가변적 시간 밀도 블록

을 추출하기 위해 motif 기반 그래프 합성곱을 가진 새로운 모델을 제안하였다. 또한 어텐션 메카니즘으로 시간 영역의 전역적 의존성을 추출하기 위해 비지역적 블록을 사용했다. Xu[25]는 스켈레톤 기반 행동인식을 위해 앙상블 신경망을 제안하였다. Base-Net으로서 residual 구조를 가지는 1차원 CNN을 설계하였다. 전체에서 지역으로, 포커싱에서 모션으로 다양한 특징을 추출하기 위해 Base-Net을 기반으로 4개의 서브넷을 설계하였다. 첫 서브넷은 전체 스켈레톤에서 시간과 공간 특징을 고려하는 2-스트림의 Entirety Net이다. 두 번째 서브넷은 세부적인 공간과 시간 특징을 추출할 수 있는 Body-part Net이다. 세 번째 서브넷은 Attention Net으로 채널별로 어텐션 메카니즘이 중요한 프레임과 특징 채널을 학습 할 수 있다. 네 번째 서브넷으로 Frame-difference Net은 모션 특징을 고려한다. 이 4개의 Net은 마지막으로 앙상블을 수행하여 최종 인식을 수행하였다. 스켈레톤기반 행동인식은 스켈레톤의 복잡한 시공간 변동성을 가지고 있어 도전적인 분야이다. 그래서 Xie[26]는 복잡한 변동성을 완화시키기 위해 시간 후 공간 재측정(temporal-then-spatial recalibration) 방법을 제안하여 시간 어텐션 재측정 모듈(TARM; Temporal Attention Recalibration Module)과 시공간 합성곱 모듈(STCM; Spatio-Temporal Convolution Module)로 구성된 MANs(Memory Attention Networks)를 설계하였다. 특히 TARM은 스켈레톤 시퀀스에서 시간 어텐션을 재측정하기 위해 새로운 attention learning network가 사용된 residual learning 모듈이 배치되었다. STCM은 어텐션 재측정된 스켈레톤 관절 시퀀스를 이미지로 처리하고 CNN을 스켈레톤 데이터의 시공간 정보를 더 모델링하도록 한다. 이 두 모듈은 엔드투엔드 방식으로 학습될 수 있는 단독 네트워크 구조를 생성하였다. Li[27]는 엔드투엔드 방식의 합성곱 동시발생 특징학습 프레임워크를 제안하였다. 각 관절의 데이터 점 수준의 정보가 독립적으로 인코딩되고 공간과 시간영역에서 의미론적 표현으로 합쳐진다. 또한 우수한 관절 동시발생 특징을 학습할 수 있는 전역정 공간 병합 방법을 소개하였다. 게다가 시간적 차이뿐만 아니라 스켈레톤 좌표가 2개 스트림 모형으로 결합한다.

어텐션 기반 행동인식에 대한 연구가 진행되어 왔다. Kamel[28]은 뎁스영상과 자세 데이터로부터 합성곱 신경망을 이용해서 행동인식 방법을 제안하였다. 행동특징을 위해 2가지 입력데이터가 사용되었다. 첫 번째 입력은 연속적인 뎁스 맵이 쌓여진 뎁스 모션 이미지이고 두 번째는 시간에 따라 사람 관절의 모션을 표현하는 제안된 방법인 움직이는 관절이다. 정확한 행동인식을 위해 특징 추출을 최대화하기 위해 3개의 CNN 채널들이 다른 입력으로 학습이 되었다. 처음 채널은 뎁스 모션 이미지이고, 두 번째 채널은 뎁스 모션 이미지와 움직이는 관절 둘 다이다. 그리고 세 번째 채널은 움직이는 관절만 학습되었다. 3개의 채널을 가진 CNN으로

부터 생성된 행동 예측들은 마지막 행동분류를 위해 혼합되었다. 정확한 행동의 점수를 최대화시키기 위해 여러 가지 혼합 스코어 연산을 제안하였다. 비디오는 매우 고차원이고 기존 RNN이 복잡한 행동 정보를 추출하기 어려운 다양한 스케일의 풍부한 사람 역학을 포함하고 있다. 그래서 Du[29]는 새로운 RSTAN(Recurrent Spatial-Temporal Attention Network)를 제안하였다. RNN의 모든 시간 단위 예측에 대해 전역적 비디오 문맥으로부터 중요한 특징들을 적응적으로 식별하는 시공간 어텐션 메커니즘을 소개하였다. 기존의 LSTM을 새로운 시공간 어텐션 모듈로 강화시켰다. 각 시간 단계에서 모듈은 현재 단계에 예측과 매우 관련있는 밀집된 시공간 행동 특징을 샘플링된 모든 비디오로부터 자동으로 학습한다. 또한 보임(appearance)과 모션 LSTM을 통일된 프레임워크로 결합하기 위해 어텐션에서 기인된 보임 모션 혼합 전략을 설계하여 2-스트림의 시공간 어텐션 모듈을 가진 LSTM을 엔드투엔드 방식으로 학습시킬 수 있다. 그리고 행위자 주변의 중요 행동 영역에 집중하도록 하는 RSTAN에 대해 행위자 어텐션 정규화를 적용하였다. 행동인식은 컴퓨터 비전에서 수 십년 동안 많은 관심을 가져왔었다. 다른 행동의 시공간 변화를 모델링 하기 위해 분별적인 시공간 특징을 추출하는 것이 중요하다. Song[30]은 스켈레톤으로부터 행동인식을 위한 분별적인 시공간 특징을 찾아내고 검출하기 위해 시공간 어텐션 모델을 제안하였다. LSTM 유닛을 가진 RNN기반 모델을 설계하였다. 학습된 모델은 각 입력 프레임에서 스켈레톤의 분별적인 관절에 선택적으로 집중할 수 있고 다른 프레임의 출력에 다른 수준의 어텐션을 줄 수 있다. 효율적인 학습을 위해 정규화된 교차 엔트로피 손실과 관절 학습 전략을 제안하였다. 게다가 시간 어텐션기반으로 행동검출을 위한 행동 시간 제안을 생성하는 방법을 개발하였다. 3차원 스켈레톤 시퀀스를 이용한 행동인식이 속도와 강인함으로 평판을 얻게 되었다. 최근 제안된 CNN기반 방법도 시공간 특징을 학습하는데 좋은 성능을 보여주었다. 그럼에도 성능에 잠재적으로 제한시키는 2가지 문제가 있다. 첫째로 기존 스켈레톤 표현은 고정된 순서의 연결된 관절이 생성된다. 일치하는 의미론적 의미가 불분명하고 관절간에 구조적 정보가 손실된다. 둘째로 기존 모델은 유용한 관절에 집중하는 능력이 없었다. 어텐션 메카니즘은 다른 관절이 정확한 인식을 위해 비균등적으로 공헌했기 때문에 스켈레톤 기반 행동인식에서 중요했다. Yang[31]은 뎁스 우선의 트리 순서를 가진 스켈레톤 표현을 스켈레톤 이미지의 의미론적 의미를 강화하고 관련된 구조적 정보를 더 잘 보존하도록 다시 설계했다. 그리고 자동으로 시공간 주요 단계에 집중하고 신뢰할 수 없는 관절 예측을 걸러내는 일반적인 2-가지 어센션 구조를 제안하였다. 제안한 일반 구조를 기반으로 정제된 가지 구조를 가지는 전역적 긴 시퀀스 어텐션 네트워크를 설계하였다. 게다가 커널의 시공간 종횡비를 조절하고 더 나은 장기 의존성을 추출하기

위해 입력으로 서브-이미지 시퀀스를 취하는 서브-시퀀스 어텐션 네트워크(SSAN; Sub-Sequence Attention Network)를 제안하였다. 2-가지 어텐션 구조는 SSAN과 결합하여 더 개선되었다. Sang[32]은 공간특징에서 행동 관련 지역적 시각 정보를 추출하기 위한 recurrent region attention cell을 제안하였다. 비디오의 시간적 순차 특성에 따라, recurrent region attention cell을 기반으로, RRA(Recurrent Region Attention) 모델을 제안하였다. RRA에서 recurrent region attention cell이 비디오의 시간적 순서를 따라 반복함에 따라 RRA의 어텐션 성능이 점진적으로 개선된다. 둘째로 전체 비디오 시퀀스에서 더 중요한 프레임을 강조하는 VFA(Video Frame Attention) 모델을 제안하였다. 그리고 엔드투엔드 방식의 학습 가능한 TAMNet(Two-level Attention Model based video action recognition Network)를 제안하였다. 오래 지속되거나 유사한 행동은 부족한 특징 시퀀스 추출을 야기하므로 인식률을 떨어뜨린다. Yu[33]는 시공간 처리성능을 개선하기 위해 단일 대상과 상호 행동 모두를 위한 3D-CNN과 LSTM기반의 새로운 분별적인 딥 모델을 제안했다. 이 모델은 RGB와 뎁스영상의 실시간 특징 혼합 방법을 사용하였다. 지역적 혼합물의 구성을 통해 더 대표적인 특징 시퀀스가 얻어져 유사한 행동의 분별 성능이 강화된다. 그리고 실시간으로 다른 가중치를 할당함으로서 개별 프레임에 집중하는 개선된 어텐션 메카니즘을 소개하였다. 또한 최고의 성능을 가진 파라미터를 얻기 위해 대체 최적화 전략(alternating optimization strategy)을 제안하였다.

물체 기반 행동인식에 대한 연구가 진행되어 왔다. Moore[34]는 이미지기반, 객체기반, 행동기반 정보를 비디오로부터 측정하여 행동과 객체를 인식하는 프레임워크를 소개하였다. 은닉 마르코프 모델들이 손 동작을 분류하기 위해 물체 컨텍스트와 결합된다. 또한 검출된 행동과 함께 저수준 추출된 물체 특징을 평가함으로서 알려지지 않은 객체의 클래스를 구분하기 위해 베이지안 방법을 사용했다. 대부분 제안된 방법들은 행동과 물체를 별개로 인식하였다. 그러나 행동은 물체와 관련이 있기 때문에 보상적으로 행동과 물체를 인식하는게 중요하다. Saitou[35]는 행동과 물체의 관계를 계층적 모델로 표현하고 비전을 통해 머리와 손의 움직임을 추적하였다. 위치와 방향의 행동특징들이 추출되고 Dynamic Bayesian Network에 입력하여 행동을 대략적으로 분류하였다. 그리고 행동과 관련된 물체가 개념적인 모델을 사용하여 정제된다. Gu[36]는 행동인식 성능을 개선하기 위해 물체/행동 상관관계와 행동 시퀀스의 제한에 대해서 행동역학뿐만 아니라 컨텍스트 제한도 모델링하는 계층적 확률 모델기반 프레임워크를 제안하였다. 행동/물체의 상관관계를 고려함으로써, 감지하기 어렵거나 인식하기 어려운 행동들도 모션 특징만을 사용해서 인식이 가능해졌다. 반면에 행동 시퀀스 제한으로 인식 정확도가

더 개선될 수 있다. 제안한 방법으로 우선 은닉 마르코프 모델을 사용하여 행동의 역학을 모델링하고 저수준 행동인식을 위한 물체 제한을 모델링하기 위해 베이지안 네트워크를 사용했다. 그렇게 해서 고수준 HMM이 생성되고 베이지안 모델로부터 결정을 정제하는 시퀀스 제한이 모델링된다. 수동적인 행동의 경우 행위자의 자세가 유사하여 분별력이 떨어져 분류에 어려움이 있다. 이 경우 인식률은 쥐고 있는 물체의 디테일에 의존한다. Rosenfeld[37]는 행동-물체의 정확한 위치를 얻어서 인식률을 개선하고 결과적으로 행위자-물체 상호작용과 함께 물체 모양의 디테일을 추출하였다. 연속적인 단계에서 의미론적 분할과 문맥적 특징을 결합한 Coarse-to-fine 방법을 적용하였다.

앙상블 기반 행동인식에 대한 연구가 진행되어 왔다. 뎁스 카메라로부터 스켈레톤 기반 행동인식을 위해, 비슷한 모션을 가진 물체 관련 행동을 구분하는 것은 어려운 일이다. 다른 이용가능한 비디오 스트림(RGB, 적외선, 뎁스)은 추가적인 단서를 제공 한다. Boissiere[38]는 스켈레톤과 적외선 데이터를 결합한 모듈 네트워크를 제안하였다. 사전학습 된 2차원 CNN은 스켈레톤 데이터로부터 특징을 추출하는 포즈 모듈로서 사용된다. 사전학습 된 3차원 CNN은 비디오로부터 시각 특징을 추출하는 적외선 모듈로서 사용된다. 이 두 특징벡터는 다층 퍼셉트론을 이용해 혼합된다. 2차원 스켈레톤 좌표는 적외선 비디오에서 대상주변의 관심영역을 잘라내는데 사용된다. 적외선 영상은 빛에 덜 민감하고 어둠에서 더 쓸모 있다. Liu[39]는 3차원 스켈레톤과 RGB이미지간의 멀티모달 혼합을 기반으로 하는 행동인식을 고려했다. 그는 3차원 스켈레톤 시퀀스와 중간의 단일 프레임을 입력으로 하는 신경망을 설계했다. 셀프 어텐션 모듈과 스켈레톤 어텐션 모듈을 사용하였고 BI-LSTM을 통해 스켈레톤 시퀀스로부터 시간 특징이 추출되었다. 그리고 공간 특징과 시간 특징이 특징 혼합 네트워크를 통해 결합된다.

제2절 행동인식 데이터셋

행동인식은 시퀀스 데이터를 이용해 특정 대상이 행하고 있는 행동을 자동으로 판별하는 것이다. 여기에서 사용되는 시퀀스 데이터는 단일 데이터가 시간 축을 따라 여러 개 쌓여 연속된 움직임 정보를 포함하는데 효율적인 데이터 형식이다. 예를 들면 비디오 데이터는 특정 너비와 높이를 가지는 2차원 이미지가 시간 축을 따라 여러 장 쌓여서 연속된 움직임 정보를 애니메이션으로 나타낼 수 있다. 또한, 사람의 스켈레톤 좌표들도 시퀀스 데이터 형식으로 시간 축을 따라 쌓여서 움직임 정보를 나타낼 수 있다.

MSR-Action3D(Microsoft Research-Action 3D)는 뎁스이미지와 스켈레톤으로 구성되어 있다. 게임 콘솔과의 상호작용 관련 20가지 행동들을 포함하고 있고, 7명의 사람으로부터 3번씩 데이터를 취득하였다. 행동을 수행하는 동안 행위자는 카메라를 응시하였고 한 손만 사용하는 행동의 경우 되도록 오른쪽 손을 사용하길 권하였다. 뎁스영상은 15 frame/s의 속도로 취득되었고, 크기는 640×480이다. 데이터는 총 23,707프레임으로 4020개 샘플로 구성되어 있다[40].

CAD-60(Cornell Activity Datasets)은 컬러이미지, 뎁스이미지, 스켈레톤으로 구성되어 있다. 물 마시기, 이 닦기 등의 12가지의 일상 행동으로 구성되어 있고 장소를 바꿔가며 3~4개의 일반적 행동들을 취득하였다. 남자 2명과 여자 2명으로부터 취득하였고 주방, 사무실, 욕실, 거실, 침실을 배경으로 획득되었다. 하나의 행동에 대해 대략 45초 동안 수집하였고 기본 행동 지침만 제공된 후로 자세한 행동은 개인마다 다양하다. 카메라로부터 행동이 가려지지 않도록 하였고 키넥트 v1으로 촬영되었다[41].

RGBD-HuDaAct(Human Daily Activity)는 컬러이미지와 뎁스이미지로 구성되어 있다. 연구실 환경에서 키넥트 v1으로 획득되었고 카메라의 위치와 방향의 변화가 적다. 사람과 카메라의 거리는 약 3미터에서 촬영되었고, 컬러이미지와 뎁스이미지 모두 640×480 해상도로 획득되었다. 컬러이미지는 23비트 RGB값이고 뎁스이미지는 16비트 정수이다. 30 frame/s의 속도로 취득되었고 컬러이미지와 뎁스이미지는 칼리브레이션이 적용되었다. 사람의 일상 행동 12가지를 30명의 학생들로부터 획득하였다. 14개 세션으로 학생마다 2-4번 행동을 수행하였다. 30~150초 분량의 비디오를 1189개 수집하였다[42].

MSRDailyActivity3D(Microsoft Research Daily Activity 3D)는 컬러이미지, 뎁스이미지, 스켈레톤으로 구성되어 있다. 거실환경에서 16가지 일상 행동들을 수집하였고 전체 행동 샘플 개수는 320개이다. 한 사람이 한 행동에 대해 두 가지 자세로 수행을 하였고 키넥트 v1을 이용해 획득되었다[43].

Act4^{2}는 컬러이미지와 뎁스이미지로 구성되어 있다. 14개의 일상 행동을 실내환경에서 다양한 카메라 위치에서 촬영되었다. 이를 위해 일반 거실에서 다른 높이와 다양한 각도의 4개의 키넥트를 사용하였다. 행동은 24명으로부터 여러 차례 획득되었다[44].

CAD-120(Cornell Activity Datasets)는 컬러이미지, 뎁스이미지, 스켈레톤으로 구성되어있다. 120개 행동 시퀀스로 구성되어 있고 4명의 사람이 일련의 행동을 각 3번씩 수행을 하였다. 카메라는 사람이 정면을 보지 않더라도 화면에 들어 올 수 있게 설치하였고 종종 자신 몸에 가려짐이 있다[45].

3D Action Pairs는 컬러이미지, 뎁스이미지, 스켈레톤으로 구성되어 있다. 6가지 상반되는 행동 페어를 정의하여 각 행동마다 3번씩 10명의 다른 사람이 수행하였다[46].

Multiview 3D Event는 컬러이미지, 뎁스이미지, 스켈레톤으로 구성되어 있다. 다른 방향에서 키넥트 3대가 동시에 데이터를 수집한다. 실내환경에서 8명이 행동을 하였고 사람마다 한 행동을 20번씩하면서 11가지의 다른 물체와 스타일로 반복하였다. 3815개의 비디오 시퀀스로 구성되어 있으며 각 행동은 평균적으로 477개 시퀀스를 가지고 있다[47].

Online RGB+D Action은 컬러이미지, 뎁스이미지, 스켈레톤으로 구성되어 있다. 거실에서 일상적인 행동 7가지를 수행하였으며, 주로 사람이 사물을 사용하는 행동들을 수행하였다. 매 프레임마다 사물을 경계박스로 손수 라벨링을 수행하였고, 키넥트 센서를 이용하였다. 첫 번째는 동일한 환경에서 16명의 사람으로부터 매 행동마다 2번씩 수행을 하였고, 두 번째는 다른 환경에서 새로운 8명의 사람으로부터 수집하였고 세 번째는 실시간 행동인식을 위한 분할되지 않은 행동들을 수집하였다[48].

Northwestern-UCLA(University of California Los Angeles)는 컬러이미지, 뎁스이미지, 스켈레톤으로 구성되어 있다. 3개의 키넥트로부터 연속 촬영되었으며 10개의 행동 범주를 포함하고 있다. 각 행동은 10명의 사람이 수행하였고 여러 방향의 카메라로부터 촬영되었다[49].

UWA(University of Western Australia)3D Multiview는 컬러이미지, 뎁스이미지, 스켈레톤으로 구성되어 있다. 대상들의 크기와 카메라 방향이 다양하고 모든 행동을 연속적으로 수행하였다. 또한 시작과 끝의 몸의 위치가 다르고 30가지 행동을 10명의 사람이 수행하였다. Cross-View 행동인식을 위해 5명은 15가지 행동에 대해 4개의 측면 방향도 촬영하였다[50].

Office Activity는 컬러이미지와 뎁스이미지로 구성되어 있다. 비디오 시퀀스 1180개로 구성이 되어 있고 사무실 환경에서 취득되었다. 3개의 키넥트 센서를 가지고 3가지 방향에서 획득하였고 변동성을 위해 두 개의 사무실에서 취득하였다. 10명의 사람이 행동을 수행하였으며 2개의 세트로 나눠서 첫 번째는 혼자만의 행동을 두 번째는 두 명의 상호행동을 취득하였다[51].

UTD-MHAD(University of Texas at Dallas-Multimodal Human Action Database)는 컬러이미지, 뎁스이미지, 스켈레톤, 관성센서데이터로 구성되어 있다. 동기화된 멀티 모달 데이터로 27가지 행동을 8명(남4, 여4)이 4번씩 반복하여 취득되었다. 키넥트 v1을 이용해 640×480 해상도의 컬러이미지, 320×240 해상도의 뎁스이미지, 20관절의 3차원 스켈레톤 좌표로 초당 30프레임 속도로 취득 되었다. 861개의 데이터 시퀀스로 구성되어 있으며 스포츠 행동, 손제스처, 일상행동, 트레이닝 운동의 행동을 포함한다. 카메라는 사람으로부터 3m 거리에서 취득되었고 관성센서는 오른쪽 손목 또는 허벅지에 착용하였다[52].

UWA(University of Western Australia)3D Multiview II는 컬러이미지, 뎁스이미지, 스켈레톤으로 구성되어 있다. 키넥트를 이용해 연구실에서 취득되었고 4가지 카메라 방향에서 사람마다 4번씩 동일한 행동을 수행하였다. 10명의 사람이 30가지 행동을 수행하였고 사람마다 30가지 행동들을 카메라 방향을 바꿔가며 4번씩 연속으로 수행하였다[53].

NTU RGB+D는 컬러이미지, 뎁스이미지, 스켈레톤, 적외선 이미지로 구성되어 있다. 56880개 RGB+D의 비디오 샘플들로 구성되어 있으며, 실내환경의 80가지의 카메라 방향에서 취득되었다. 10세부터 35세까지 40명이 일상, 상호, 건강 관련 행동을 60가지 수행하였고 키넥트 v2를 이용해 취득되었다[54]. 표 2-1은 행동인식에 대한 공개 데이터셋을 보여준다. C(Color)는 컬러이미지, D(Depth)는 뎁스이미지, 3DJoints는 스켈레톤, IR(Infrared Ray)은 적외선 이미지를 나타낸다.

노인과 성인의 행동특성분석을 위해 행동인식 데이터는 도메인이 구분되어야 하고 실질적인 노인의 일상 행동들을 정의해서 현실적 분석이 가능해야 한다. 또

한, 휴먼케어 로봇상황에서 다양한 뷰와 거리, 배경에서 데이터가 취득되어야 한다.

표 2-1. 행동인식에 대한 공개 데이터셋

데이터셋	샘플	클래스	대상	방향	센서	데이터타입
MSR-Action3D	567	20	10	1	N/A	D+3DJoints
CAD-60	60	12	4	-	Kinect v1	C+D+3DJoints
RGBD-HuDaAct	1189	13	30	1	Kinect v1	C+D
MSRDailyActivity3D	320	16	10	1	Kinect v1	C+D+3DJoints
Act4[2]	6844	14	24	4	Kinect v1	C+D
CAD-120	120	10+10	4	-	Kinect v1	C+D+3DJoints
3D Action Pairs	360	12	10	1	Kinect v1	C+D+3DJoints
Mltiview 3D Event	3815	8	8	3	Kinect v1	C+D+3DJoints
Online RGB+D Action	336	7	24	1	Kinect v1	C+D+3DJoints
Northwestern-UCLA	1475	10	10	3	Kinect v1	C+D+3DJoints
UWA3D Multiview	~900	30	10	1	Kinect v1	C+D+3DJoints
Office Activit	1180	20	10	3	Kinect v1	C+D
UTD-MHAD	861	27	8	1	Kinect v1	C+D+3DJoints+ID
UWA3D Multiview II	1075	30	10	5	Kinect v1	C+D+3DJoints
NTU RGB+D	56880	60	40	80	Kinect v1	C+D+3DJoints+IR

제3절 행동인식 방법

최근 딥러닝이 발달하면서 기존의 복잡한 문제들도 컴퓨터가 자동으로 수행할 수 있는 기틀이 마련되었다. 딥러닝은 기존 신경망에서 은닉층을 깊게 쌓고 역전파 알고리즘을 이용해 학습시키는 것으로 비선형 문제도 잘 해결하는 특징을 보인다. 행동인식에도 그러한 딥러닝 기술을 적용한 연구가 진행되고 있다[55,56].

이미지의 다운 샘플링은 이미지에서 미세한 정보를 잃어버리기 때문에 좋은 결과를 낳지 못하고 고해상도 이미지를 그대로 사용하면 추론 시간을 증가시키는 문제가 있다. 그래서 Karpathy[57]는 비디오 분류를 위해 병렬로 실행되는 두 개의 스트림을 융합하는 것을 제안하였다. 병렬로 실행되는 2개의 인코더를 작게 하여 더 적은 파라미터로 단순화시켰고, 하나의 인코더는 저해상도를 취하고 다른 하나는 고해상도를 처리하여 마지막 단의 전연결레이어에서 결합된다.

융합 방식은 짧은 비디오에 대해 잘 작동하지만 긴 비디오를 분류하는 데에는 많은 프레임을 계산하고 많은 것을 기억해야 하므로 더 어렵다. Ng[58]은 긴 비디오를 분류하기 위해 2가지를 제안하였다. 첫 번째는 합성곱 특징들을 시간축에 대해 맥스풀링을 사용하고 두 번째는 다양한 길이의 비디오를 처리하도록 합성곱 특징들을 LSTM(Long Short Term Memory)으로 연결하였다.

비디오에서 물체의 모션은 수행하는 행동에 대해 좋은 정보를 주고 그러한 물체의 모션은 옵티컬 플로우(optical flow)를 이용해 측정할 수 있다. Simonyan[59]은 이미지와 옵티컬 플로우로부터 2개의 스트림을 이용하는 행동인식 방법을 제안하였다. 하나의 스트림은 개별적 프레임을 입력하고 다른 하나의 스트림은 여러 프레임들을 취해서 옵티컬 플로우를 계산한다. 그리고 각각의 CNN에 입력하여 마지막 단에서 둘의 스코어를 결합한다.

2차원 합성곱은 2차원 데이터를 취해서 2차원 결과를 출력하는 반면 3차원 합성곱은 세 방향으로 합성곱 연산을 하므로 3차원 데이터를 입력으로 하여 3차원의 결과를 출력할 수 있다. Tran[60]는 비디오 행동인식을 위해 3차원 합성곱 연산을 기반으로 하는 3차원 합성곱 신경망 구조를 제안하였다. 그 구조는 8개의 합성곱 레이어와 2개의 전연결레이어로 구성되어 있다.

Wang[61]은 수행된 행동을 분류하기 위해 바디파트의 궤적(trajectory)를 사용하였다. 이 연구는 비디오에서 궤적을 추출한 후 피셔벡터(Fisher)의 손수 추출한

특징과 CNN기반의 심층 학습된 특징을 마지막 레이어에서 결합하였다.

Yang[62]은 비디오 분류를 위해 4개의 모델을 가지는 멀티 모달 결합을 제안하였다. 이 4가지 모델을 3차원 합성곱 특징, 2차원 옵티컬 플로우, 3차원 옵티컬 플로우, 2차원 합성곱 특징이다. 결합 방식은 부스팅(bootsting) 메커니즘을 사용하였다.

주의집중(attention) 메커니즘은 인식 활동에 대한 영역에 집중하는 방법으로 다른 영역보다 특정영역에 더 많은 가중치를 주는 방법이다. 이 가중치는 데이터로부터 학습되며 일반적으로 소프트(soft)와 하드(hard) 방식으로 나뉜다. 소프트는 결정적이고 하드는 확률적인 방식이다. Shama[63]는 비디오를 분류하는데 주의집중 메커니즘을 적용하였다. 합성곱 특징맵과 위치 가중치를 3개의 LSTM에 직렬 입력하여 위치 확률을 구한다. 이 주의집중은 정확도를 개선해줄 뿐만 아니라 예측을 가시화하는 방법도 제공한다.

행동인식 데이터는 보통 RGB비디오와 스켈레톤 등으로 구성되어 여러 특성의 정보를 제공해주는 대신 크고 복잡하다. 이러한 다양한 특성의 입력 데이터를 효과적으로 활용한 연구가 적고 단일 분류모델만으로 특징을 다양하게 추출하기 어렵다. 또한, 데이터가 큰 만큼 불필요한 정보도 많이 포함된다.

제4절 행동특성분석 방법

딥러닝은 데이터의 특성에 따라 여러 가지 기본 신경망들로 세분화되어 발달되었다. 이미지의 공간 특징을 효율적으로 추출하기 위한 심층 합성곱 신경망[64], 시간 특징을 효율적으로 추출하기 위한 순환 신경망(Recurrent Neural Network)[65], 분자 모형 또는 소셜 네트워크와 같이 관계 및 상호작용과 같은 추상적인 개념을 다루기에 적합한 그래프 신경망(Graph Neural Network)[66] 등이 있다. 이렇듯 데이터 특성에 따라 신경망을 적절하게 사용해야 최적의 성능을 얻을 수 있기 때문에 데이터의 특성을 분석하는 것은 매우 중요하다.

Horst[67]는 2개의 압력판과 10개의 적외선 카메라를 이용해 걸음걸이 동안의 관절각도와 지면반력을 측정하였고 딥러닝을 이용해 개인식별을 진행하였다. 그리고 학습된 모델을 LRP(Layer-wise Relevance Propagation)방법을 이용해 역으로 분해하여 어떤 변수가 어느 시점에서 기여됐는지 가시화하고 그 결과를 개인의 걸음걸이 특성으로 해석하였다.

Notthoff[68]는 50세 이상 사람들을 대상으로 신체활동의 개인차이가 상당히 있는 이유를 알기 위해 인구통계적인 특성, 건강, 심리적 요인을 고려한 문헌들을 시스템적으로 검토하여 분석하였다.

Johannsen[69]은 젊은 사람과 나이든 사람뿐만 아니라 90대 이상 사람들의 신체활동을 비교하기 위해 14일 넘게 총 에너지 소비량(total energy expenditure)과 휴식 대사율(resting metabolic rate)를 측정하여 통계적 분석을 하였다.

Harris[70]는 젊은 사람과 늙은 사람의 일상 비운동 활동의 관계를 명확히 하기 위해 지방제외체중, 신체 자세를 10일 넘게 측정하였고 에너지 소비를 가늠하여 통계적으로 비교분석하였다.

Goble[71]은 노인과 성인의 복잡한 양손동작(bimanual task) 수행능력을 비교하기 위해 기능적 자기공명 영상(functional magnetic resonance imaging)에서 활성화 되는 부분을 분석하였다.

노인과 성인의 행동특성분석에 있어 주로 데이터의 통계적 수치의 비교로 연구가 이루어졌다. 노인과 성인의 행동특성을 인지과학적으로 분석한 연구가 적다.

제3장 제안하는 행동인식 방법

제1절 기존 행동인식을 위한 기법

1. 스켈레톤 시퀀스의 이미지화

스켈레톤은 센서 데이터를 기반으로 사람의 신체 골격을 좌표점들로 재구성한 것으로 사람의 움직임을 효율적으로 저장할 수 있는 데이터 형식이다. 행동인식을 위해서 이미지처럼 한 순간의 스켈레톤은 사람의 행동 정보를 모두 포함하지 못하기 때문에 비디오처럼 여러 순간의 스켈레톤을 시간순으로 시퀀스를 만들어 사용한다. 이 시퀀스 데이터를 효과적으로 분석하기 위해 공간정보뿐만 아니라 시간정보도 중요하기 때문에 그에 맞는 정보를 효과적으로 추출하기 위한 변환 방법들이 연구되었다. 여기서는 스켈레톤 시퀀스를 하나의 컬러이미지로 변환하는 방법인 PEI(Pose Evolution Image)를 소개한다. 먼저, 관절은 신체가 접혀질 수 있는 중심축으로 보통 사람은 접힐 수 있는 관절이 제한적이기 때문에 적은 데이터만으로도 사람의 골격을 나타낼 수 있다. 키넥트 v1은 20개 관절로 사람의 골격을 나타내는 반면 키넥트 v2은 25개 관절로 사람의 골격을 나타낸다. 관절이 많음으로서 사람의 세부적인 골격변화도 감지할 수 있지만 불필요하게 많은 관절은 사람의 신체제한을 고려하지 못해서 잘못된 골격을 검출할 수도 있다. 스켈레톤은 이 관절들의 묶음이며 3차원 스켈레톤은 사람의 신체 골격의 관절들이 3차원 좌표로 표현된다. 행동은 시간에 따라 사람이 움직이면서 신체골격도 따라 변하므로 그 스켈레톤을 일정 시간마다 계속 잡아내야 된다. 그렇게 해서 한 행동에 대해 생성된 스켈레톤 시퀀스는 3차원의 데이터 형식을 가진다. 이 3차원 데이터를 2차원 이미지로 변환하는 방법은 3차원 좌표를 RGB공간에 그대로 투영하는 것이다. 그림 3.1은 스켈레톤을 RGB공간에 투영하여 이미지화하는 원리를 보여준다. 그림 3.2는 스켈레톤 시퀀스의 이미지화 과정을 나타낸다. 스켈레톤 시퀀스가 3차원 데이터로서 (J×D×T)로 표현되면 J는 신체골격을 나타내는 관절들의 개수이고 D는 관절을 나타내는 좌표의 차원수이다. T는 시간적 차원으로 시간에 따른 스켈레톤 프레임 수이다. 스켈레톤 시퀀스를 이미지로 변환하기 위해서 관절좌표의 차원(D)을 시간적 차원(T)과 치환한다. 관절좌표의 차원수(D)가 3개일 경우 치환 과정을 거치면 한

장의 컬러이미지(J×T×3) 형태를 가진다. 이 컬러이미지를 채널별로 정규화를 수행하고 이미지 크기를 선형 변환하면 스켈레톤 이미지가 생성된다. 사전학습 된 2D-CNN은 이미지 인식을 위해 주로 RGB의 3채널을 입력받도록 설계되어 있기때문에 스켈레톤 시퀀스를 PEI로 변환함으로써 사전학습 된 2D-CNN에 바로 사용할 수 있다. 또한, 스켈레톤 시퀀스를 PEI로 변환함으로써 2차원 필터만으로 시공간 특징을 모두 고려할 수 있다. 그림 3.3은 PEI의 전·후 특징추출 비교를 나타내고 식 3-1은 채널별 정규화 수식을 나타낸다. 이 이미지화 방법을 타입1(Type1)로 정의한다[72].

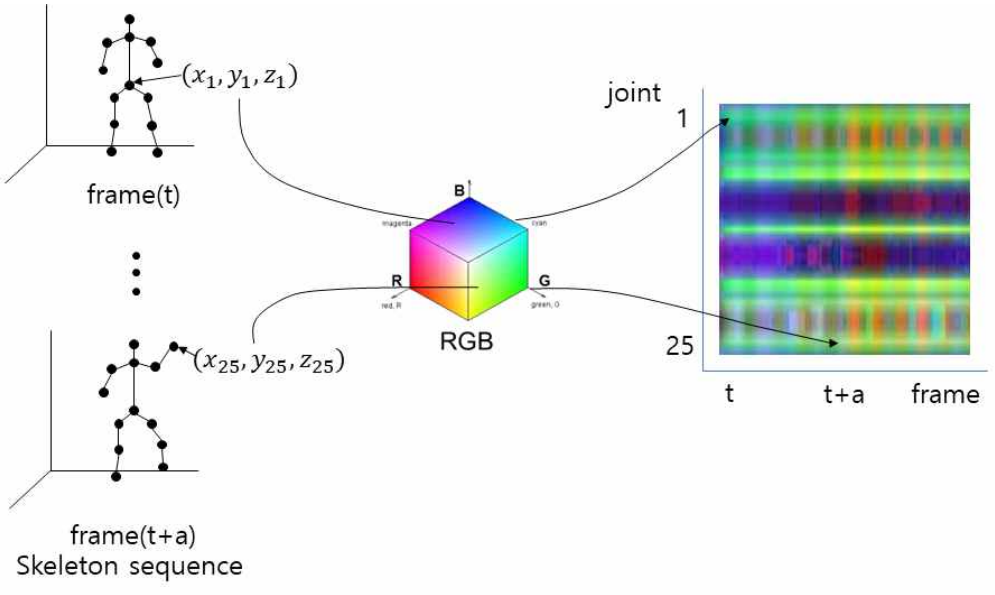


그림 3.1 스켈레톤을 RGB공간에 투영하여 이미지화하는 원리

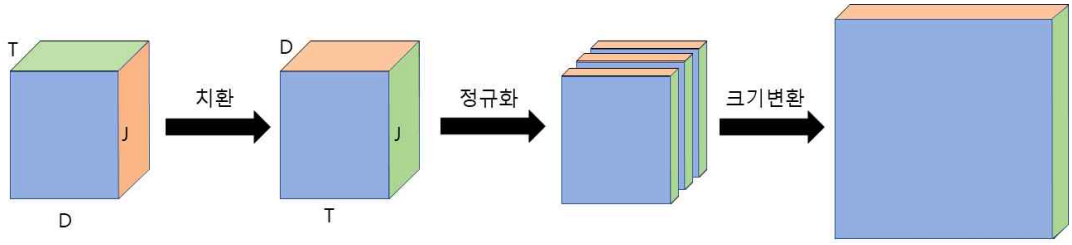


그림 3.2 스켈레톤 시퀀스의 이미지화 과정

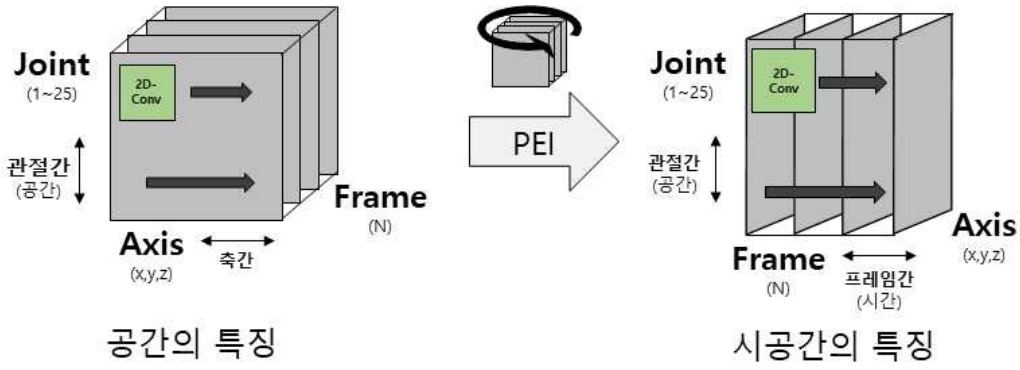


그림 3.3 PEI의 전 · 후 특징추출 비교

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3-1}$$

앞서 스켈레톤 시퀀스의 이미지화 방법은 또한 스켈레톤 데이터에 변화를 주어 다양한 이미지를 얻을 수 있다. 3차원의 원래 스켈레톤 좌표를 골반라인을 기준으로 회전시켜 회전된 스켈레톤의 좌표들을 얻고 그 좌표들을 이용해 앞서의 방법으로 이미지화를 수행한다. 이 이미지화 방법을 타입2(Type2)로 정의한다. 그림 3.4는 스켈레톤의 회전을 보여준다.

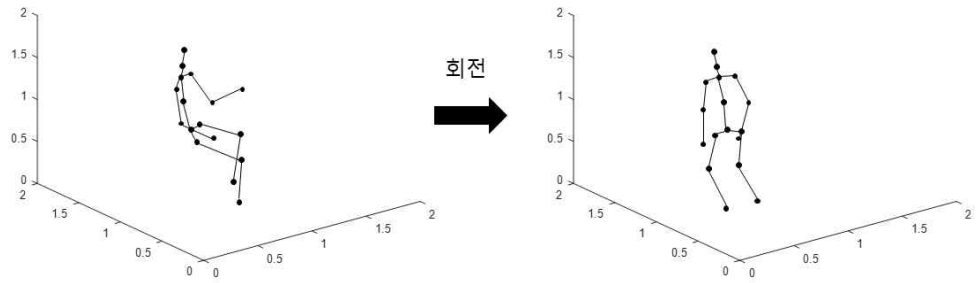


그림 3.4 스켈레톤의 회전

3차원의 원래 스켈레톤 좌표에서 이웃하는 두 좌표 사이에 새로운 관절들을 삽입하여 앞서의 방법으로 이미지화를 수행한다. 이 이미지화 방법을 타입3(Type3)

로 정의한다. 그림 3.5는 스켈레톤의 관절삽입을 보여준다.

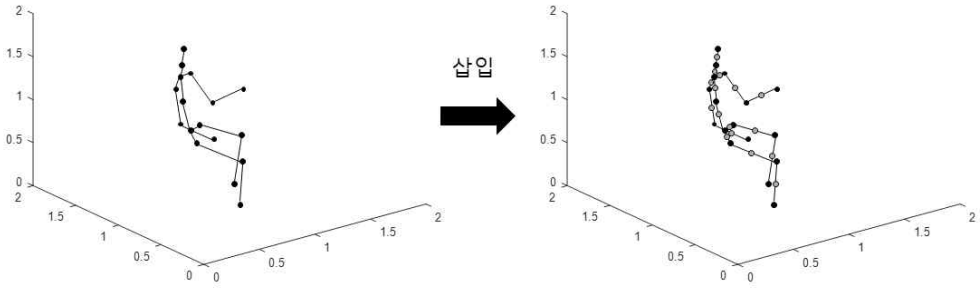


그림 3.5 스켈레톤의 관절삽입

마지막으로 3차원의 원래 스켈레톤 좌표에서 회전과 삽입을 모두 적용하여 앞서의 방법으로 이미지화를 수행하고 이를 타입4(Type4)로 정의한다[72].

2. 심층 신경망

머신러닝에서 신경망은 인간 뇌의 구조 및 동작을 모방하여 디지털 데이터를 인식하는 방법으로 인간이 직접 논리를 설계하는 것이 아니라 컴퓨터가 스스로 논리를 만든다. 뉴런은 신경계의 구조적 기능적 단위인 신경세포이며 정보 전달의 기본 단위로서 신경 신호를 만들어 내고 이 신호를 몸의 한 부분에서 다른 부분으로 전달한다. 뉴런은 수상돌기, 세포체, 축색돌기로 이루어져 있고, 수상돌기는 외부 신호를 입력받고 세포체, 축삭돌기를 지나 다음 뉴런에 신호를 전달한다. 이 과정은 입력된 신호가 역치 값을 넘어서야 다음 뉴런에 전달된다. 이런 신경세포의 기능을 컴퓨터상에서 모델링하여 인공지능 기술에 활용하는 것을 신경망이라고 한다. 신경망에서 하나의 뉴런은 노드로서 정의가 되고 역치 값은 활성함수로서 정의가 된다. 수많은 뉴런 중에서 서로 강하게 연결된 뉴런도 있고 연결이 되지 않은 뉴런도 있는데 이는 노드간의 가중치로서 정의가 된다. 초창기 신경망은 노드만으로 구성된 층들이 얕은 구조를 가진다. 이러한 단순한 구조는 단순한 문제에 대해서 작동하고 문제가 복잡해지면 학습이 되지 못한 문제가 있다. 신경망은 레이어 층을 깊게 쌓고 다양한 기능의 레이어들을 추가하여 이제는 복잡한 문제도 학습할 수 있는 모델들이 개발되었다. 신경망의 특성에 따라 여러 기본적인 신경망이 있다. 이미지 분석에 유리한 CNN(Convolutional Neural Network)[73], 시퀀스 데이터 분석에 유리한 RNN(Recurrent Neural Network)[74], 계층 데이터 분석에 유리한 GNN(Graph Neural Network)[66] 등이 있다.

기존 영상처리 방식은 전문가의 지식을 바탕으로 특징추출을 위한 신호처리 과정을 구현하고 추출된 특징을 분류기로 분류를 하였는데, CNN은 데이터로부터 스스로 특징추출을 하고 분류하는 알고리즘이다. 이미지의 특징을 효과적으로 추출하기 위해 2차원 공간을 따라 통과하는 합성곱 필터를 적용한 합성곱 레이어와 이동과 크기의 변화에 안정적인 서브샘플링 레이어, 분류를 위한 전연결 레이어, 소프트맥스 등으로 구성되어 있다. 식 3-2는 2차원 합성곱 연산을 보여준다.

$$(f*g)(i,j) = \sum_{x=0}^{h-1} \sum_{y=0}^{w-1} f(x,y)g(i-x,j-y) \tag{3-2}$$

그림 3.6과 같이, 합성곱 연산은 이미지에서 필터 크기 부분만 떼어내어 필터와 동위상의 각 원소마다 곱을 한 후 모두 더해서 하나의 결과 값을 출력한다. 이러한 과정은 필터가 이미지를 모두 스캔하는 동안 계속된다. 그림 3.6는 2차원 합성곱을 보여준다.

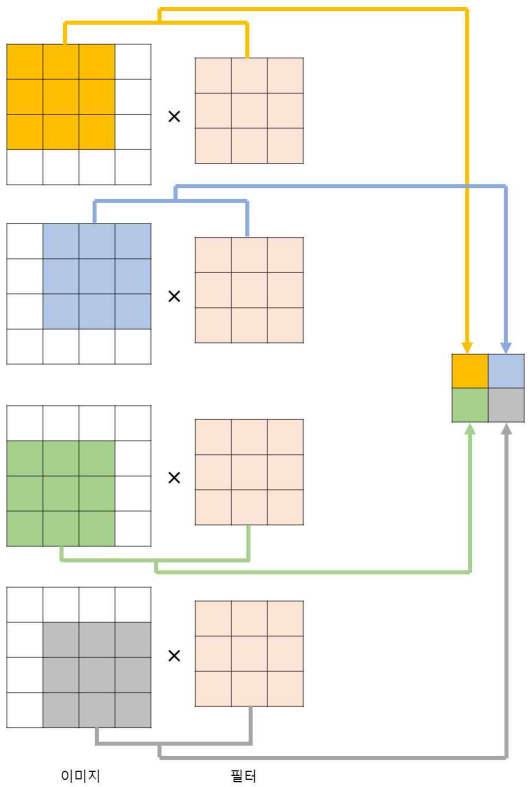


그림 3.6 2차원 합성곱

그림 3.7과 같이, 서브샘플링은 이미지에서 필터 크기 부분만 떼어내어 그 중에서 대푯값으로 데이터를 축소하는 방법이다. 주로 평균값 또는 최대 값을 사용한다. 그림 3.7은 서브샘플링을 보여준다.

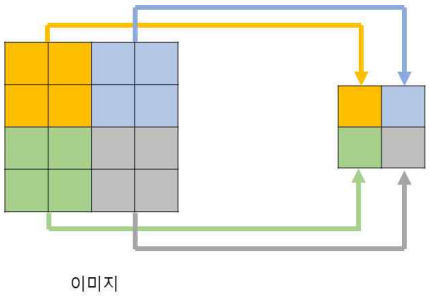


그림 3.7 서브샘플링

소프트맥스는 입력받을 값을 0과 1사이 값의 확률로 변환하는 것으로 출력 값들의 총합이 1이 되는 특성을 가진 함수이다. 식 3-3은 소프트맥스 함수를 나타낸다.

$$f(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}}, \ \text{i=1,...,K} \tag{3-3}$$

활성함수로는 시그모이드 함수, 하이퍼탄젠트 함수, ReLU(Rectified Linear Unit) 등이 있다. 그림 3.8은 시그모이드 함수, 그림 3.9는 하이퍼탄젠트 함수, 그림 3.10은 ReLU 함수를 보여준다.

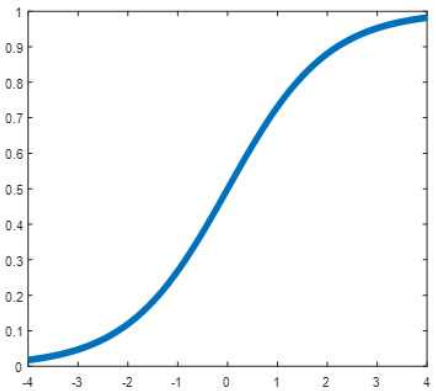


그림 3.8 시그모이드 함수

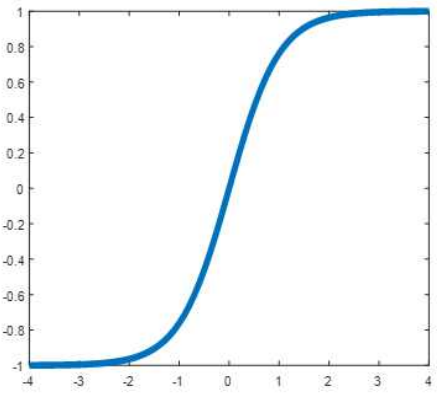


그림 3.9 하이퍼탄젠트 함수

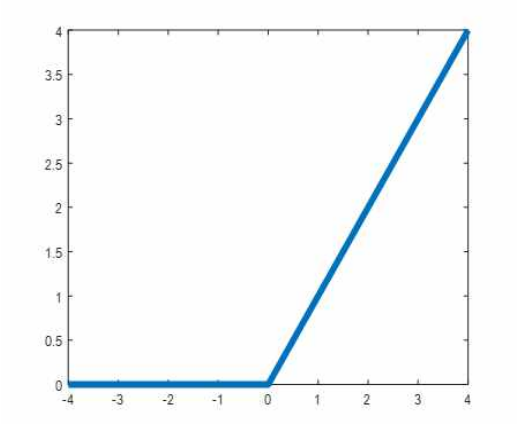


그림 3.10 ReLU

CNN은 합성곱 레이어와 서브샘플링 레이어를 반복하며 데이터의 차원을 축소시키고 말단 부분에서 전연결레이어를 통해 특징벡터로 만든 후 마지막 단에서 소프트맥스를 통과시켜 분류를 수행한다. 이외에도 과적합을 예방하기 위한 배치정규화, 드롭아웃 등이 있다. 그림 3.11은 간단한 CNN의 구조를 나타낸다. Dense 레이어는 단순 뉴런들의 연결로 구성되는 레이어이다[73].

그림 3.11 간단한 CNN의 구조

2차원 CNN은 이미지의 공간적인 특징밖에 추출을 못하지만 3차원 CNN은 공간적인 특징뿐만 아니라 시간적인 특징도 효과적으로 추출할 수 있다. 단일 이미지의 경우 2차원 공간 정보만을 포함하는 것과 달리 비디오는 여러 장의 이미지가 겹쳐진 3차원 데이터로 공간정보와 시간정보를 가지고 있기 때문에 2차원 합성곱만으로 충분한 특징을 추출하기 어렵다. 3차원 합성곱은 필터가 3차원이기 때문에 공간특징과 시간특징을 모두 추출이 가능하기 때문에 비디오와 같은 3차원 데이터에 효율적이다. 3차원 CNN은 합성곱 연산과 풀링 연산이 3차원 필터를 가지고 수행되며 일반적인 구조는 2차원 CNN과 동일하다. 식 3-4은 3차원 합성곱 수식을 보여준다.

$$(f*g)(i,j,k) = \sum_{x=0}^{h-1}\sum_{y=0}^{w-1}\sum_{z=0}^{t-1} f(x,y,z)g(i-x,j-y,k-z) \qquad (3\text{-}4)$$

그림 3.12처럼 3차원 합성곱 연산은 이미지에서 필터 크기 부분만 떼어내어 필터와 동위상의 각 원소마다 곱을 한 후 모두 더해서 하나의 결과값을 출력한다. 2차원 합성곱과 다른 점으로 3차원 합성곱은 시간축까지 고려한다. 이러한 과정은 필터가 이미지를 모두 스캔하는 동안 계속 된다. 마찬가지로 서브샘플링도 이미지에서 필터 크기 부분만 떼어내어 그 중에서 대푯값으로 데이터를 축소하는 방법이다. 2차원 서브샘플링과 다른 점으로 3차원 서브샘플링은 시간축까지 고려한다. 그림 3.12은 3차원 합성곱을 보여준다. 3차원 CNN은 2차원 CNN과 비슷하게 2차원 합성곱 및 2차원 서브샘플링 대신에 3차원 합성곱 및 3차원 서브샘플링을 사용하고 활성함수, 배치정규화, 드롭아웃 등을 이용해 설계 된다[75].

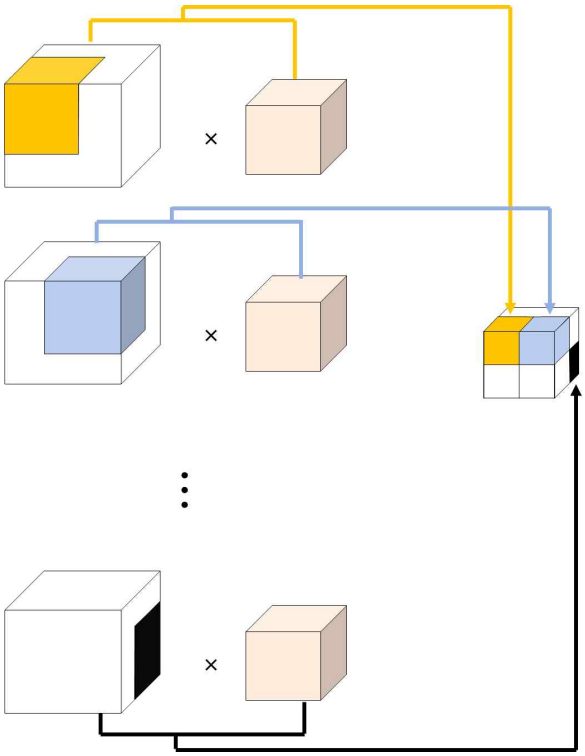


그림 3.12 3차원 합성곱

LSTM(Long Short Time Memory)은 순서가 있는 데이터를 입력받고 그 안에서 규칙을 찾아내어 출력을 얻는다. 순서가 있는 데이터는 시간에 따라 변화하는 데이터로서 자연어, 노래, 날씨 등이 있다. 또한, 행동 데이터도 시간에 따라 움직임 정보를 가지는 시퀀스 데이터이기 때문에 LSTM을 통해 효과적으로 행동을 분석할 수 있다. LSTM은 RNN(Recurrent Neural Network)의 진보된 신경망으로 일반 신경망과 달리 출력을 입력에 다시 넣는 되먹임 구조를 특징으로 갖는다. 그림 3.13은 LSTM의 되먹임 구조를 보여주고 그림 3.14는 LSTM의 파이프라인을 나타낸다.

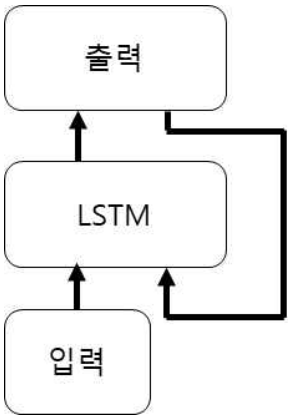


그림 3.13 LSTM의 되먹임 구조

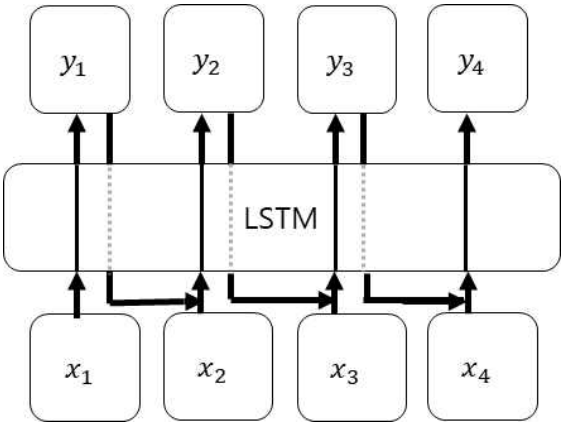


그림 3.14 LSTM의 파이프라인

LSTM은 입력과 출력의 길이가 고정되지 않고 가변적인 특징을 가져서 입력과 출력의 개수를 변경하여 다양한 응용이 이루어지고 있다. LSTM은 RNN의 진보된 모델로 입력 데이터가 길어질수록 학습이 잘되지 않는 장기의존성 문제를 셀 상태를 추가하고 내부 구조를 더 복잡하게 만들어 개선되었다. 그림 3.15는 LSTM의 계산 흐름도를 보여주고 수식 3-5부터 수식 3-10은 LSTM 계산 과정을 보여준다.

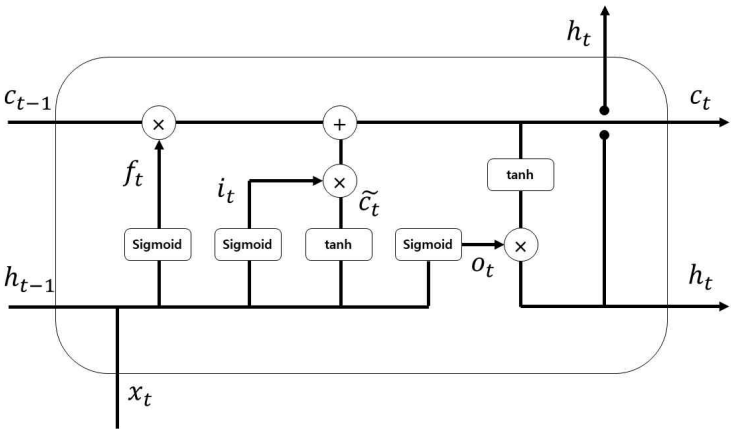


그림 3.15 LSTM의 계산 흐름도

$$i_t = sigmoid(x_t w^i + h_{t-1} u^i) \tag{3-5}$$

$$f_t = sigmoid(x_t w^f + h_{t-1} u^f) \tag{3-6}$$

$$o_t = sigmoid(x_t w^o + h_{t-1} u^o) \tag{3-7}$$

$$\tilde{c_t} = \tanh(x_t w^{\tilde{c}} + h_{t-1} u^{\tilde{c}}) \tag{3-8}$$

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c_t} \tag{3-9}$$

$$h_t = \tanh(c_t) \times o_t \tag{3-10}$$

w와 u는 입력과 출력에 곱해지는 가중치이다. i는 입력(input), f는 망각(forget), o는 출력(output) 게이트를 통과한 결과를 나타낸다[74].

제2절 비디오와 스켈레톤의 앙상블기반 행동인식

행동인식 데이터는 RGB비디오만으로 구성되기도 하지만 일반적으로 스켈레톤 시퀀스도 포함하는 경우가 많다. RGB비디오는 동영상으로 RGB이미지가 일정 간격의 시간차이를 두고 여러 장 연속 촬영한 데이터이다. 동영상은 연속 촬영된 이미지들을 연속으로 보여줌으로 마치 실제 움직이는 것처럼 보여주는 것이다. 스켈레톤은 센서 데이터로부터 사람의 골격정보를 추출한 것으로 머리, 어깨, 손, 발 등의 관절좌표로 구성되어 있고 매 프레임마다 정의되어 스켈레톤 시퀀스를 형성한다. RGB비디오는 이미지의 해상도에 따라 데이터의 크기 차이가 크고 일반적으로 다른 데이터보다 용량이 수십 배 크지만 그만큼 주변 물체와 상황을 포함하여 다양한 정보가 포함되어 있다. 반면에 스켈레톤 데이터는 관절 좌표정보만을 포함하기 때문에 데이터가 작으며 오직 사람의 골격구조 정보만을 포함하고 있다. 행동인식에 있어 중요한 정보는 사람의 골격 움직임이지만 비슷한 행동의 경우 골격정보만으로는 부족하여 주변 상황에서 판단해야하는 경우도 있다. 이 두 종류의 데이터가 가지고 있는 특성이 다르기 때문에 이 두 데이터를 적절히 앙상블함으로써 더 좋은 시너지효과를 낸다.

RGB비디오는 2차원의 이미지가 시간축을 따라 겹겹이 쌓여 있어 3차원 구조를 가지기 때문에 2차원 CNN에 적용하기에는 어려움이 있다. 그래서 2차원 이미지의 프레임마다 2D-CNN을 이용해 공간정보에 대한 특징벡터를 생성하고 그 특징벡터를 다시 LSTM에 입력으로 넣어 시간정보에 대한 특징을 추출하여 분류한다. LSTM은 출력을 구하기 위해 이전 레이어들의 수치를 참조하기 때문에 시퀀스 데이터 처리에 효율적인 신경망이다. 특징추출기로서 2D-CNN은 새로 설계한 모델부터 사전학습 된 모델까지 다양하게 적용할 수 있다. 사전학습 된 모델은 세계 뛰어난 석학들의 연구에 기반 되어 성능이 검증된 모델로 수십에서 수백의 레이어를 정의해야 하는 수고를 줄이고 성능도 보장받을 수 있기 때문에 유용하다. 사전학습 된 모델들로는 가볍고 비교적 얕은 모델부터 무겁고 깊은 모델까지 다양하며 AlexNet[73], GoogLeNet[76], ResNet[77], DenseNet[78] 등이 있다. 2D-CNN-LSTM의 CNN모델로는 GoogLeNet을 사용한다. GoogLeNet은 9개의 인셉션 모듈로 구성된 사전학습 된 모델이다. 인셉션모듈은 신경망의 깊이를 효율적으로 깊게 하기 위해 다양한 크기의 합성곱 연산을 병렬적으로 수행한 후 결합하는 레이어들의 묶음이

다. 이 인셉션 모듈은 1×1 크기의 합성곱 연산을 통해 특징맵의 개수를 줄이면서 최소 단위의 미세한 특징도 고려되고 또한 3×3, 5×5 크기의 다양한 합성곱이 포함되어 있다. 병렬적 풀링을 통해 이전 레이어의 내용을 요약하고 마지막에 모든 병렬적 결과들을 연결한다. 그림 3.16은 인셉션 모듈의 구조를 보여주고, 그림 3.17은 RGB비디오 입력의 2D-CNN 특징추출기와 LSTM을 보여주고 있다.

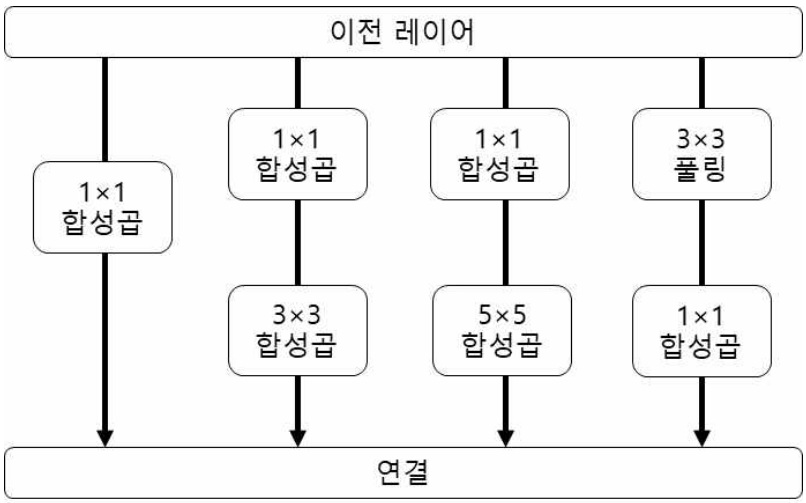


그림 3.16 인셉션 모듈의 구조

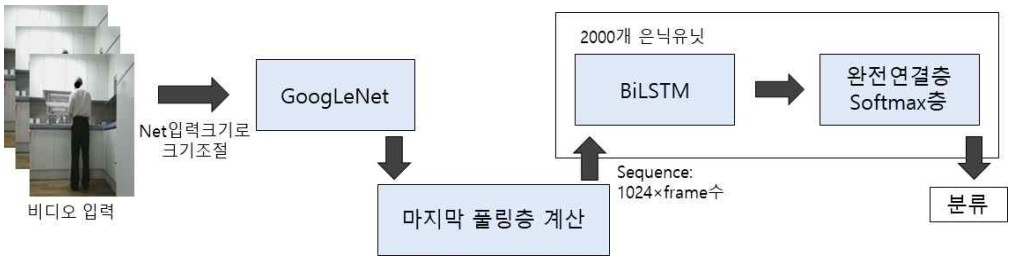


그림 3.17 RGB비디오 입력의 2D-CNN 특징추출기와 LSTM

RGB비디오는 2차원의 이미지가 시간축을 따라 겹겹이 쌓여 있어 3차원 구조를 가지기 때문에 2차원 CNN에 적용이 어려워 3차원 CNN을 이용해야 한다. 3차원 CNN은 3차원 필터를 가지기 때문에 따로 시퀀스에 대응하지 않더라도 필터가 3차원이기 때문에 시공간 정보가 모두 고려된다. 합성곱 연산과 서브샘플링이 3차원 필터를 가지고 그 밖의 구성은 2D-CNN과 동일하며 사전학습 된 모델도 2D-CNN과 같게

설계하여 좋은 성능을 낼 수 있다. 사전학습 된 모델로는 C3D[60], GoogLeNet을 기반으로 하는 I3D[79], ResNet을 기반으로 하는 ResNet3D(R3D)[80] 등이 있다. ResNet은 레이어가 깊어질수록 기울기가 매우 작아지거나 커지는 문제, 레이어가 깊을수록 성능이 더 나빠지는 문제를 해결하기 위해 이전 레이어의 입력특징을 재사용하는 스킵 커넥션(skip connection)을 적용하였다. ResNet은 5개의 블록을 만들어서 하나는 입력으로, 나머지는 순서대로 중복하여 적층한다. Res블록을 2개씩 중첩하여 설계할 경우 R3D-18모델이다[77]. 그림 3.18은 스킵 커넥션을 보여주고 그림 3.19은 R3D-18 구조를 보여준다. 그리고 표 3-1은 R3D-18의 구조를 나타내고 그림 3.20은 RGB비디오 입력의 3D-CNN을 보여준다.

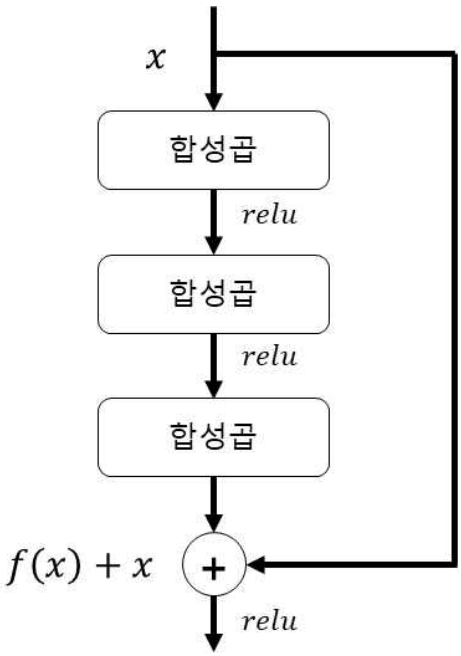


그림 3.18 스킵 커넥션

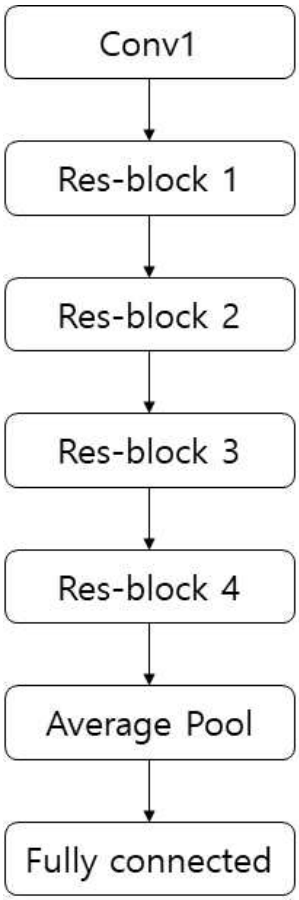


그림 3.19 R3D-18 구조

표 3-1. R3D-18의 구조

레이어 이름	R3D-18		출력
합성곱1	5×5×5, 64, Stride(2,2,2)		32×38×32×64
최대값 풀링	1×3×3, Stride(1,2,2)		32×19×16×64
Res 블록1	1×3×3, 64, Stride(1,1,1)	×2	32×19×16×64
	1×3×3, 64, Stride(1,1,1)		
Res 블록2	1×3×3, 128, Stride(1,2,2)	×2	32×10×8×128
	1×3×3, 128, Stride(1,1,1)		
Res 블록3	3×3×3, 256, Stride(1,2,2)	×2	32×5×4×256
	3×3×3, 256, Stride(1,1,1)		
Res 블록4	3×3×3, 512, Stride(1,2,2)	×2	32×3×2×512
	3×3×3, 512, Stride(1,1,1)		
평균 풀링	32×3×2		1×1×1×512
전연결	512×2		2

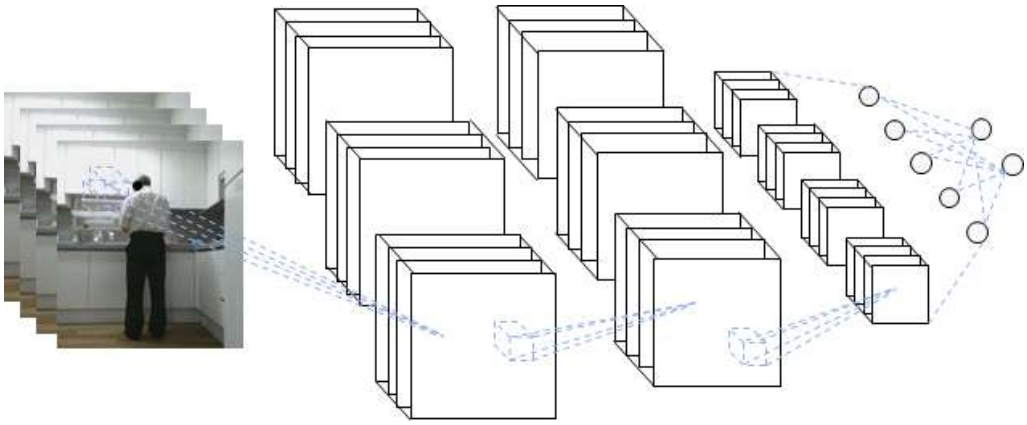


그림 3.20 RGB비디오 입력의 3D-CNN

사람은 관절을 회전축으로 움직이기 때문에 관절을 잘 설계하면 사람을 스켈레톤 데이터로 모델링 할 수 있다. 예로 스켈레톤 데이터 취득에 널리 사용되는 키넥트 v2는 25개의 관절로 사람을 모델링 하고 있다. 키넥트 v2는 관절점들을 3차원 좌표로 획득하고 그 정의된 관절들은 표 3-2과 같다. 그림 3.21는 키넥트 v2 스켈레톤의 관절 위치를 나타낸다[81].

표 3-2. 키넥트 v2의 스켈레톤 관절명칭

ID	Joint
1	Spine base
2	Spine emid
3	Neck
4	Head
5	Left shoulder
6	Left elbow
7	Left wrist
8	Left hand
9	Right shoulder
10	Right elbow
11	Right wrist
12	Right hand
13	Left hip
14	Left knee
15	Left ankle
16	Left foot
17	Right hip
18	Right knee
19	Right ankle
20	Right foot
21	Spine shoulder
22	Left hand tip
23	Left thumb
24	Right hand tip
25	Right thumb

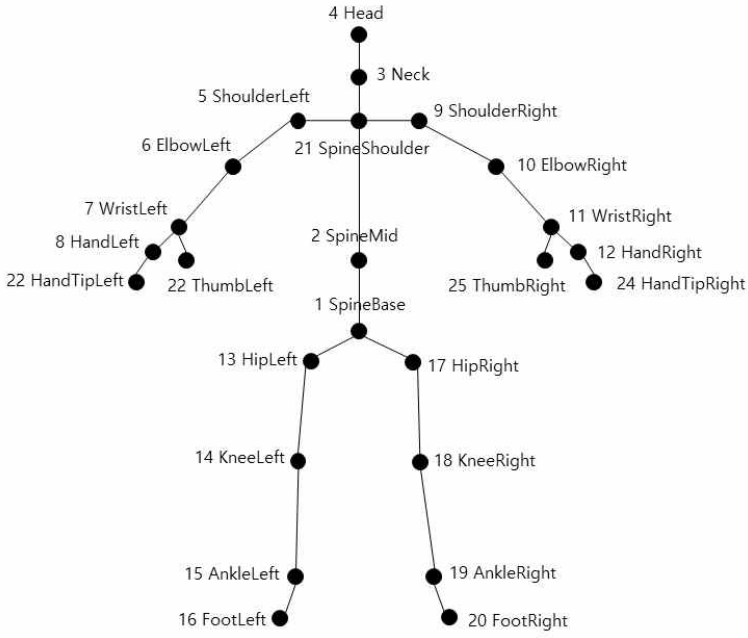


그림 3.21 키넥트 v2의 스켈레톤 관절위치

스켈레톤 데이터가 3차원 좌표상에서 25개의 관절이 정의될 때, 시간축이 더해지면 이 스켈레톤 데이터도 3차원 데이터 형식을 가진다. 이 3차원 스켈레톤 데이터를 PEI방법으로 변환할 경우 2차원 이미지로 바뀌게 되고 이 2차원 이미지는 2D-CNN을 학습시켜 분류할 수 있다. 여기에서 2D-CNN은 특징추출 및 분류기로서 마찬가지로 GoogLeNet, ResNet 등과 같은 사전학습 된 모델을 사용할 수 있다. 그림 3.22은 PEI입력의 2D-CNN을 보여주고 있다. 앞서 PEI에 대해 살펴보았듯, 원래 스켈레톤 데이터에 변화를 주어 4가지 타입의 PEI를 생성 할 수 있고 타입별로 모델을 학습시켜 4가지 2D-CNN 모델들을 얻을 수 있다.

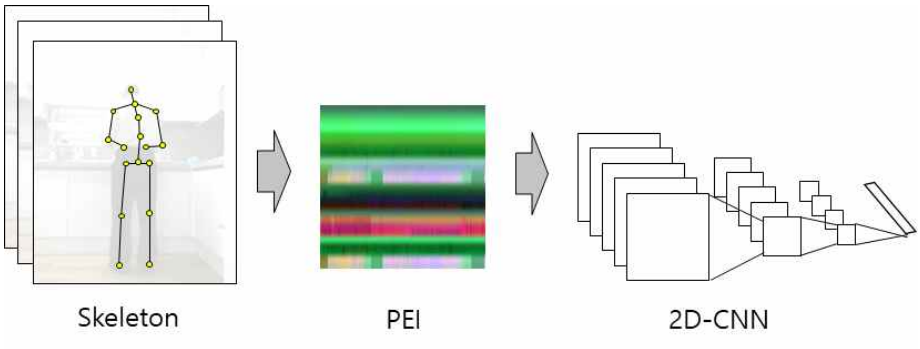


그림 3.22 PEI입력의 2D-CNN

이렇게 얻어진 PEI의 4가지 타입에 대한 2D-CNN 모델들을 앙상블하여 더 좋은 성능이 얻어진다. PEI의 4가지 타입은 앞서 살펴보았듯, 원래 스켈레톤 데이터를 기준으로 생성한 것, 스켈레톤에 회전을 주어 생성한 것, 관절점을 추가하여 생성한 것과 스켈레톤에 회전과 관절점 추가를 모두 하여 생성한 것이다. 모델의 앙상블을 위해서 PEI 타입별 모델의 출력 스코어를 모두 더하거나 곱해서 그 결과의 최대 값을 가지는 클래스로 분류한다.

신경망의 앙상블은 한 가지 목표를 가지고 개별적으로 학습된 여러 모델들의 결과를 합쳐서 더 나은 결과를 도출하는 방법이다. 개별 모델들은 서로 다른 입력 데이터에 대해 흐트러지지 않고 각자 맡은 특징에 집중한다. 또한, 개별 모델들은 서로 다른 신경망 구조를 통해 데이터의 분석 전략을 다양화시킬 수 있다. 이러한 다양한 입력과 분석전략의 모델들을 앙상블하여 더 나은 강한 분류기가 된다. 앙상블 방법으로는 신경망 모델의 출력 스코어 값들을 더하거나 곱하는 방법이 있다. 식 3-11는 출력 스코어의 덧셈을 나타내고, 식 3-12는 출력 스코어의 곱셈을 나타낸다. 그림 3.23은 신경망의 출력단에서 앙상블 방법을 나타낸다.

$$output_{plus} = \max\left(\begin{bmatrix} p_{A1} \\ p_{A2} \\ p_{A3} \end{bmatrix} + \begin{bmatrix} p_{B1} \\ p_{B2} \\ p_{B3} \end{bmatrix}\right) \tag{3-11}$$

$$output_{product} = \max\left(\begin{bmatrix} p_{A1} \\ p_{A2} \\ p_{A3} \end{bmatrix} \times \begin{bmatrix} p_{B1} \\ p_{B2} \\ p_{B3} \end{bmatrix}\right) \tag{3-12}$$

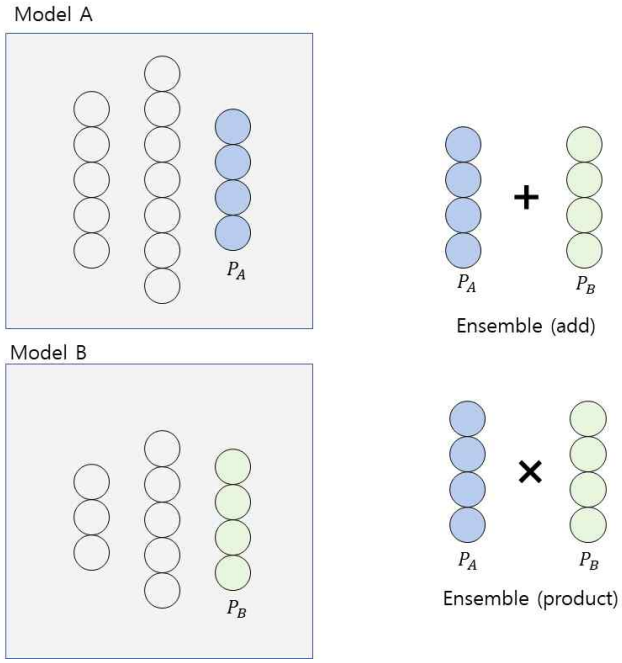


그림 3.23 신경망의 출력단에서 앙상블 방법

행동인식 데이터는 보통 RGB비디오와 스켈레톤 시퀀스의 두 종류의 데이터를 가지고 있고 그 둘 데이터의 특성이 다르기 때문에 이 두 데이터를 적절히 앙상블 하므로써 더 좋은 시너지효과를 낸다. 본 절에서 RGB비디오기반 행동인식 모델 2개와 PEI를 이용한 스켈레톤 시퀀스기반 행동인식 모델 1개를 설계하였다. 이 RGB비디오기반 행동인식 모델과 PEI-T3의 스켈레톤 시퀀스기반 행동인식 모델의 출력 스코어들을 더하거나 곱해서 그 결과의 최대값으로 최종 분류하는 방법으로 앙상블하였다. RGB-S기반의 3-스트림 앙상블 모델로서 RGB비디오가 가지고 있는 상황정보와 색상정보, 스켈레톤 시퀀스가 가지고 있는 사람의 골격정보가 최종 행동인식을 위해 각각 고려됨으로서 더 좋은 성능을 낸다. 그림 3.24은 행동인식을 위한 RGB-S기반의 3-스트림 앙상블 모델을 보여준다.

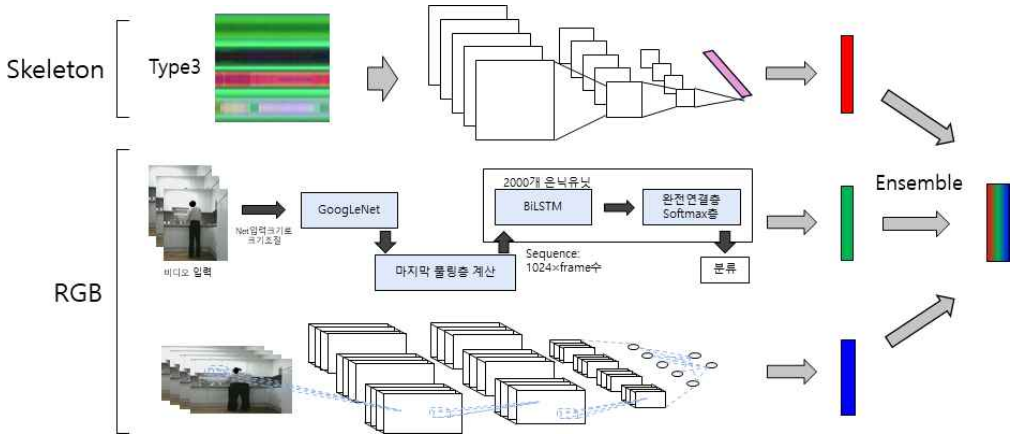


그림 3.24 행동인식을 위한 RGB-S기반의 3-스트림 앙상블 모델

제3절 Body ROI와 Hand-object ROI기반 행동인식

사람은 눈으로 들어온 빛을 통해 장면을 보고 객체를 인식하는데 이때 동시적으로 장면의 모든 곳을 보지 못하고 관심 부분 하나하나 초점을 이동시키며 인식하게 된다. 이는 목표 프로세스를 위해 불필요한 정보를 무시하고 관심목표에 좀 더 집중하여 실수를 줄이고 정확도를 높일 수 있다. 행동인식을 위한 RGB비디오 데이터는 우선 사람이 등장하고 그 행동을 수행하는 장소의 풍경과 사용하는 도구뿐만 아니라 많은 물체가 등장할 수 있다. 이때 행동인식을 위해서는 주변 풍경, 주변 물체들보다는 행동의 주체인 사람이 핵심 정보를 가지게 되기 때문에 주변의 불필요한 정보들은 제거하고 사람부분에만 관심을 두어 분석하면 더 나은 성능을 낸다. 본문에서 사람부분을 관심영역으로 설정하는 것을 Body ROI(Region Of Interest)로 지칭한다. 마찬가지로 사람의 손 영역은 사람이 행동하는데 도구를 사용하기 때문에 행동인식에 중요한 정보를 제공할 수 있다. 본문에서 손부분을 관심영역으로 설정하는 것을 Hand-object ROI로 지칭한다.

RGB비디오에서 사람의 관심영역을 추출하려면 먼저 RGB영상에서 사람을 인식하여 위치를 특정해야 한다. 이는 딥러닝기반의 RGB이미지에서 스켈레톤 정보를 추출해주는 오픈포즈(openpose)를 이용해 관절좌표들을 얻을 수 있다[82]. 오픈포즈는 RGB이미지에서 사람의 골격을 인식해 2차원 관절좌표들을 반환해주는 공개 소프트웨어이다. 오픈포즈는 사람을 25개 관절로 모델링한 스켈레톤을 제공하고 그 25개 관절들은 표 3-3와 같다. 그림 3.25은 오픈포즈의 스켈레톤 관절위치를 나타낸다.

RGB비디오에서 사람부분만 관심영역으로 지정하는 방법은 스켈레톤 데이터를 이용하여 사람부분만을 남겨두고 나머지부분은 모두 픽셀값을 0으로 하여 검은색으로 만든다. 사람부분만을 남겨두기 위해 관절좌표를 중심으로 좌우 일정 크기의 박스구간을 복사하여 동일 크기의 빈 이미지에 동일 좌표에 대입하여 넣는다. 이 과정은 모든 관절들에 대해 수행되어 사람 전신부분만 빈 이미지에 복사되게 된다. 모든 프레임에 대해 배경을 지우고 다시 비디오로 만듦으로써 Body ROI 데이터가 준비된다. 앞의 과정을 사람의 손부분에 대해서만 수행하면 Hand-object ROI 데이터가 준비된다. 그림 3.26은 스켈레톤을 이용한 RGB비디오의 Body ROI 추출 과정을 보여주고 있다. 이렇게 준비된 데이터는 배경부분이 지워진 ROI 된 RGB비

디오로 이미지들이 시간축을 따라 쌓여진 3차원 데이터이다. 이를 분류하는 방법으로 3D-CNN을 사용하였다. 그림 3.27은 Body ROI 된 RGB비디오 입력의 3D-CNN을 보여준다. 마찬가지로 그림 3.28은 스켈레톤을 이용한 RGB비디오의 Hand-object ROI 추출 과정을 보여주고 있다. 이렇게 준비된 데이터는 배경부분이 지워진 ROI 된 RGB비디오로 이미지들이 시간축을 따라 쌓여진 3차원 데이터이다. 이를 분류하는 방법으로 3D-CNN을 사용하였다. 그림 3.29은 Hand-object ROI 된 RGB비디오 입력의 3D-CNN을 보여준다.

표 3-3. 오픈포즈(openpose)의 스켈레톤 관절명칭

ID	Joint
1	Nose
2	Neck
3	Right shoulder
4	Right elbow
5	Right wrist
6	Left shoulder
7	Left elbow
8	Left wrist
9	Middle Hip
10	Right hip
11	Right knee
12	Right ankle
13	Left hip
14	Left knee
15	Left ankle
16	Right eye
17	Left eye
18	Right ear
19	Left ear
20	Left big toe
21	Left small toe
22	Left heel
23	Right big toe
24	Right small toe
25	Right heel

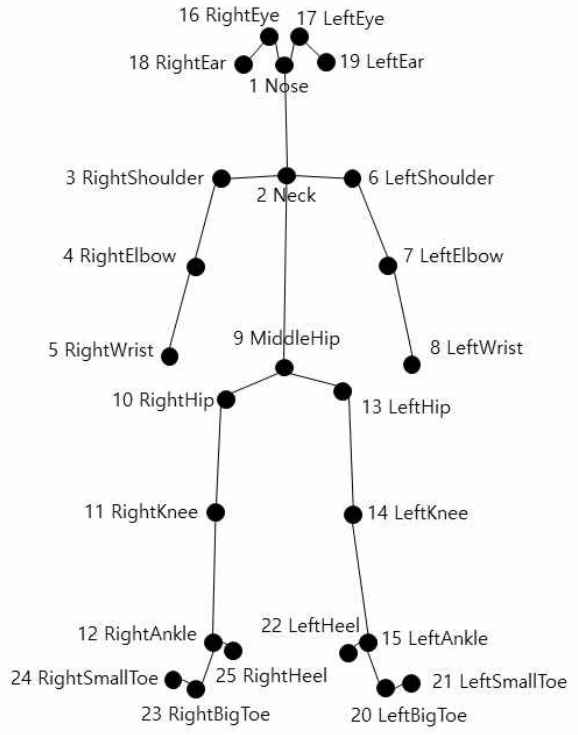


그림 3.25 오픈포즈의 스켈레톤 관절위치

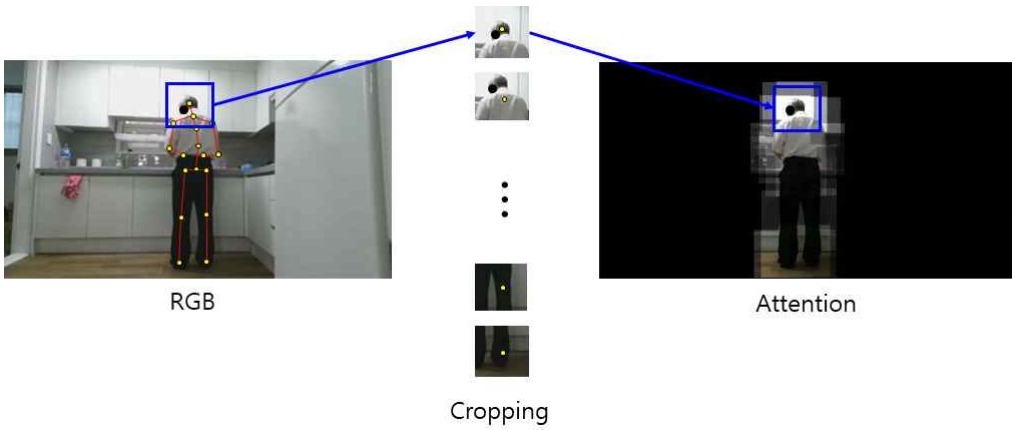


그림 3.26 스켈레톤을 이용한 RGB비디오의 Body ROI 추출 과정

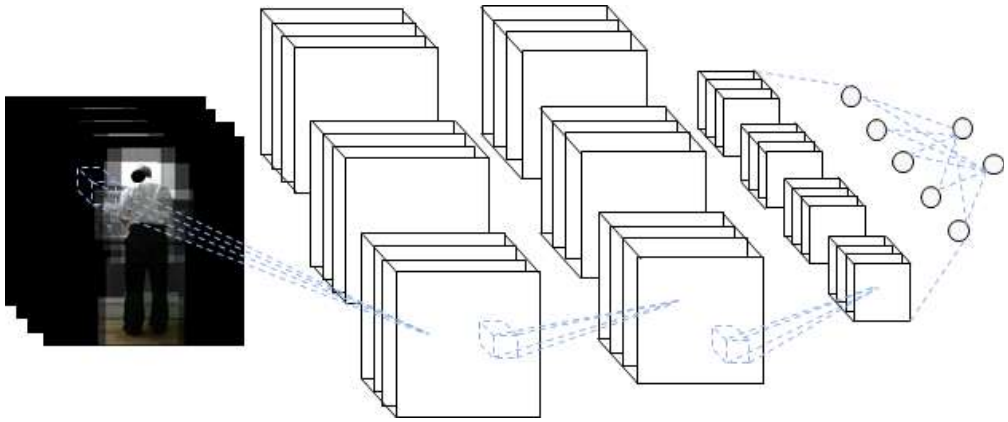


그림 3.27 Body ROI 된 RGB비디오 입력의 3D-CNN

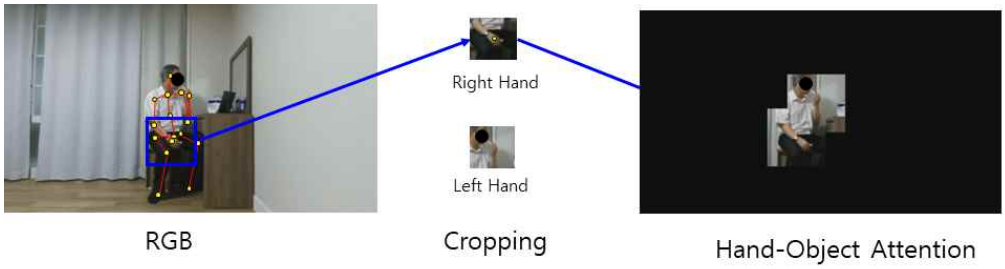


그림 3.28 스켈레톤을 이용한 RGB비디오의 Hand-object ROI 추출 과정

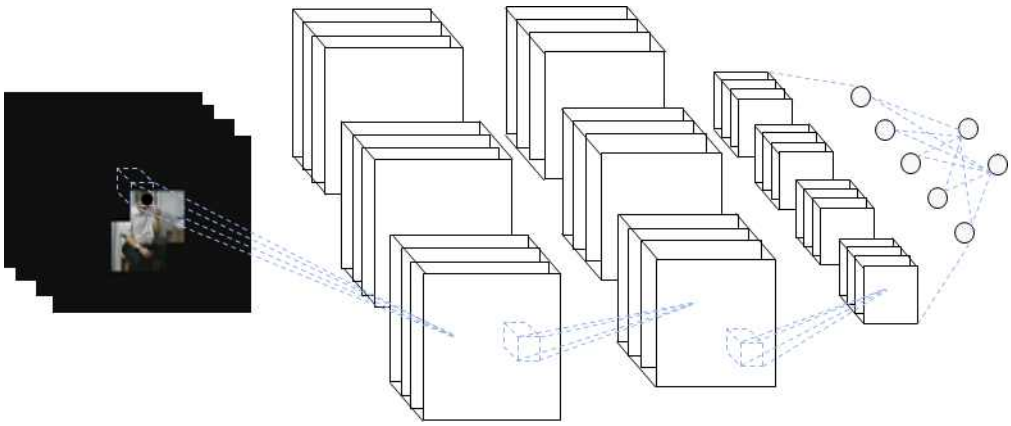


그림 3.29 Hand-object ROI 된 RGB비디오 입력의 3D-CNN

관심있는 영역만을 이용하여 3D-CNN을 새롭게 학습하므로 이미지의 전체 부분으로 학습시킬 때와는 또 다른 특징들이 추출되고 다른 분석을 하므로 이 두 입력 방식에 대한 출력 스코어 값들을 앙상블하므로써 더 좋은 시너지효과를 낸다. 그림 3.30은 행동인식을 위한 ROI기반의 4-스트림 앙상블 모델을 보여준다.

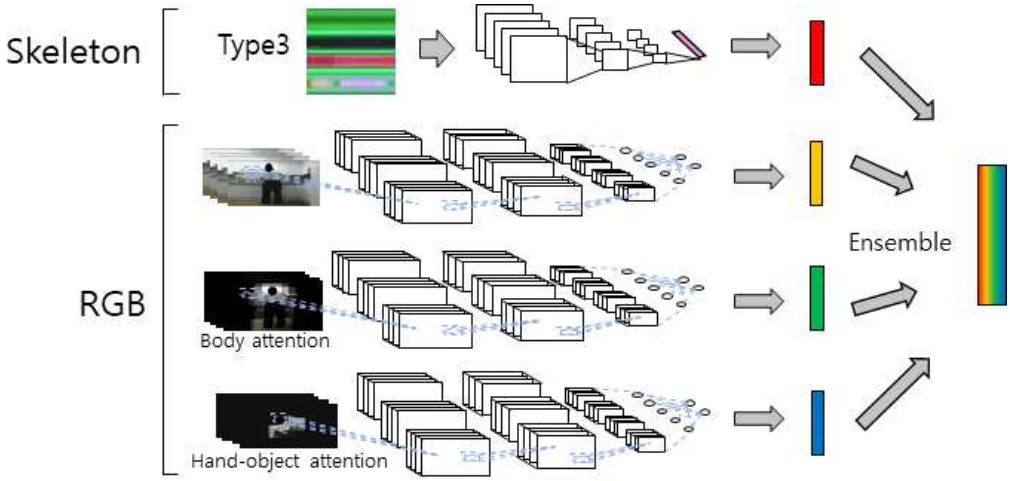


그림 3.30 행동인식을 위한 ROI기반의 4-스트림 앙상블 모델

제4절 설명 가능한 AI를 이용한 행동특성분석

의료기술이 발달하면서 사람수명이 늘어나 노인인구가 청년대비 급격하게 증가하고 있다. 그로 인해 노인의 복지, 헬스케어 등을 위한 노인연구 또한 관심이 커지고 있다. 사람은 나이가 들수록 신체적 구조가 바뀌고 신체적 기능이 떨어지므로 성인과 노인 간에 행동특성에 변화가 생긴다. 노인은 시각, 청각 및 근력이 떨어져 외부 환경을 지각하는데 더디고 반응이 성인에 비해 느린 특성을 보인다.

ETRI-Activity3D는 행동특성분석이 가능하게끔 노인과 성인의 도메인이 구분된 행동 데이터로, 표 3-4는 ETRI-Activity3D에 대한 노인과 성인의 통계치 비교를 나타낸다. 노인에 대한 프레임 길이의 평균과 분산이 성인보다 더 큰 것으로부터 노인이 성인보다 행동이 더 느린 것을 이해할 수 있다[83].

표 3-4. ETRI-Activity3D에 대한 노인과 성인의 통계치 비교

	노인	성인
프레임 길이의 평균	267	189
프레임 길이의 분산	1.189e+04	5.663e+03
모션 미분의 평균	16.79	20.29
모션 미분의 분산	1.1519e+05	4.458e+05

본 논문에서는 설명가능한 인공지능 기법을 이용하여 행동특성을 분석한다. 머신러닝은 수많은 파라미터를 가지고 오랜 시간 계산을 통해 최적의 해를 찾아내 인식함으로 인간은 기계가 왜 인식을 잘하는지 알 수가 없다. 설명 가능한 인공지능은 이러한 머신러닝의 의사 결정 과정을 인간이 이해할 수 있는 수준까지 분해하는 기술이다[84].

행동 데이터를 머신러닝의 모델에 학습시켜 자동으로 행동특성이 모델에 녹아들게 되고 그 모델의 분류과정을 도메인별 비교분석함으로서 행동특성분석이 가능하다. 이 과정은 데이터를 이용해 복잡한 계산 과정을 거쳐 나온 결과이기 때문에

인간이 놓치거나 모를 수 있는 부분까지도 분석할 수 있는 잠재력을 가지고 있다.

합성곱 신경망(Convolutional Neural Network)은 2차원 이미지를 입력 받아서 특징을 추출해 분류하는 방법으로 이미지 인식에 널리 사용되면서 합성곱 신경망의 설명가능한 AI(Artificial Intelligence)방법도 연구가 많이 진행되고 있다. CAM(Class Activation Mapping)은 출력결과에 대해 신경망 모델이 입력 이미지의 어느 부분을 보고 분류했는지를 보여준다. 기존 합성곱 신경망은 마지막 단에 전연결레이어가 위치하여 입력 이미지에 대한 공간정보가 소실되는 문제로 역추정하는데 어려움이 있었다. 그래서 전연결레이어 대신 전역 평균 풀링(Global Average Pooling) 레이어로 바꾸어 공간정보를 유지하면서 특징벡터를 구성하고 가중치를 곱해 출력을 얻는다. 그 가중치는 전역 평균 풀링 레이어 직전의 특징맵의 채널에 대한 중요도가 되어 분류에 기여되는 정보를 나타내는 히트맵을 생성하는데 사용된다[85]. Grad-CAM은 합성곱 신경망의 끝단을 변경해야하는 단점을 해결하기 위해 결정의 중요부분을 찾는데 마지막 합성곱 레이어의 기울기를 사용한다. 출력의 스코어(y^c)를 특징맵의 활성함수(A_{ij}^k)로 미분하여 기울기($\partial y^c/\partial A_{ij}^k$)를 계산하고 그 기울기는 중요도 가중치($\alpha_k^c$)를 구하기 위해 거꾸로 전파되면서 전역 평균 풀링된다. 다음 식은 Grad-CAM을 구하는 과정을 나타낸다[86].

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3-13}$$

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{3-14}$$

노인과 성인의 차이를 시각화하기 위한 방법으로 t-SNE(Stochastic Neighbor Embedding)을 사용하였다. t-SNE는 t분포를 이용해 유사도(similarity)를 측정하고 유사한 것끼리 묶어 차원을 축소하는 방법이다[87]. 스켈레톤 시퀀스를 PEI로 변환 후 ResNet-101을 학습시키고, 그 학습된 ResNet-101의 특징벡터를 추출하여 t-SNE 방법으로 2차원까지 축소하여 가시화한다. 그림 3.31은 t-SNE를 이용한 CNN 특징의 가시화 과정을 보여준다.

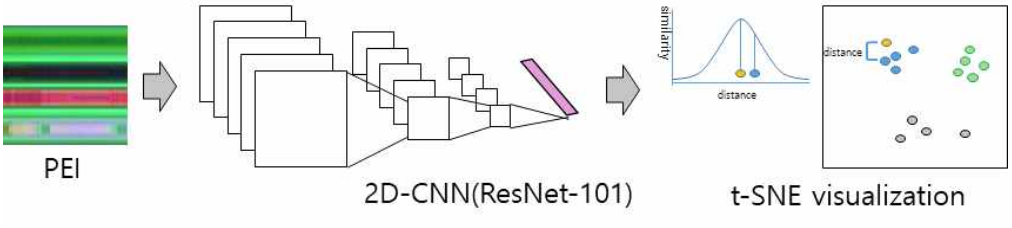


그림 3.31 t-SNE를 이용한 CNN특징의 가시화 과정

다음으로 고차원적 행동특성 분석을 위해 인지과학적인 행동특성 분석 방법을 소개한다. 인지과학적 행동특성분석을 위해 키넥트 v2로 취득한 스켈레톤 시퀀스를 PEI-Type1로 변환하고 2D-CNN을 학습시켰다. 여기서는 행동특성분석을 위해 신경망의 학습을 단일 행동에 대한 노인과 성인의 분류를 다뤘다. 학습이 완료되면 검증 데이터에 대해 2D-CNN은 입력된 데이터가 성인의 행동인지 노인의 행동인지 결과를 출력한다. 그리고 출력된 결과를 Grad-CAM(Class Activation Mapping) 방법을 통해 입력의 어떤 부분을 보면서 분류하는지를 히트맵으로 나타냈다. 히트맵은 따뜻한 색(붉은색)과 차가운 색(푸른색)으로 표시되며 따뜻한 색이 분류 결과에 주로 기여되고 있음을 의미한다. PEI-Type1에 대한 히트맵이므로 그림 3.32처럼 두 이미지를 겹쳤을 때 붉은색 부분이 그 행동으로 분류되는데 중요 모션이 된다. PEI와 히트맵 상태에서 행동적 특성을 분석하는 것이 어렵기 때문에 PEI를 스켈레톤으로 다시 바꾸어 히트맵을 스켈레톤에 겹침 출력하였다. 히트맵도 PEI의 축과 의미가 같기 때문에 x축은 프레임, y축은 관절에 해당하며 이를 바탕으로 특정 프레임과 특정 관절을 히트맵의 해당 색상으로 표현할 수 있다. 하지만 프레임마다 스켈레톤 히트맵을 출력하는 것은 또한 한눈에 행동의 끝까지 보기 어려운 문제가 있어 행동분석을 어렵게 한다. 그래서 중요한 부분인 따뜻한 색 부분의 프레임과 관절에서만 히트맵 색상을 출력하고 처음부터 끝 프레임까지 모두 중첩시켜 궤적을 가시화하였다. 이를 히트궤적으로 정의한다. 히트궤적은 전체 프레임을 한 화면에 모두 출력하기 때문에 스켈레톤 모양의 초점을 잃어버릴 가능성이 있다. 그래서 히트맵의 붉은색 부분만 출력하되 관절 부위마다 색상을 구분하여 히트된 궤적이 어느 관절인지 가독성을 개선하였다. 그림 3.33은 히트궤적의 색상 배치를 보여준다. 척추 라인은 빨간색, 오른팔은 초록색, 왼팔은 파란색, 오른 다리는 노란색, 왼 다리는 청록색으로 표시하였다. 또한 엉덩이(hip)를 중심으로 말단부위는 밝은색, 중심부위는 어두운색으로 표시하였다. 그리고 단일 데이터에 대

해 출력된 히트궤적을 RGB비디오와 비교분석하며 특성을 해석하였다. 동일 행동의 노인간 비교, 동일 행동의 노인과 성인의 비교 등을 통해 개인 행동특성뿐만 아니라 노인 행동특성도 해석할 수 있다. 그림 3.34은 행동특성분석 방법을 보여준다.

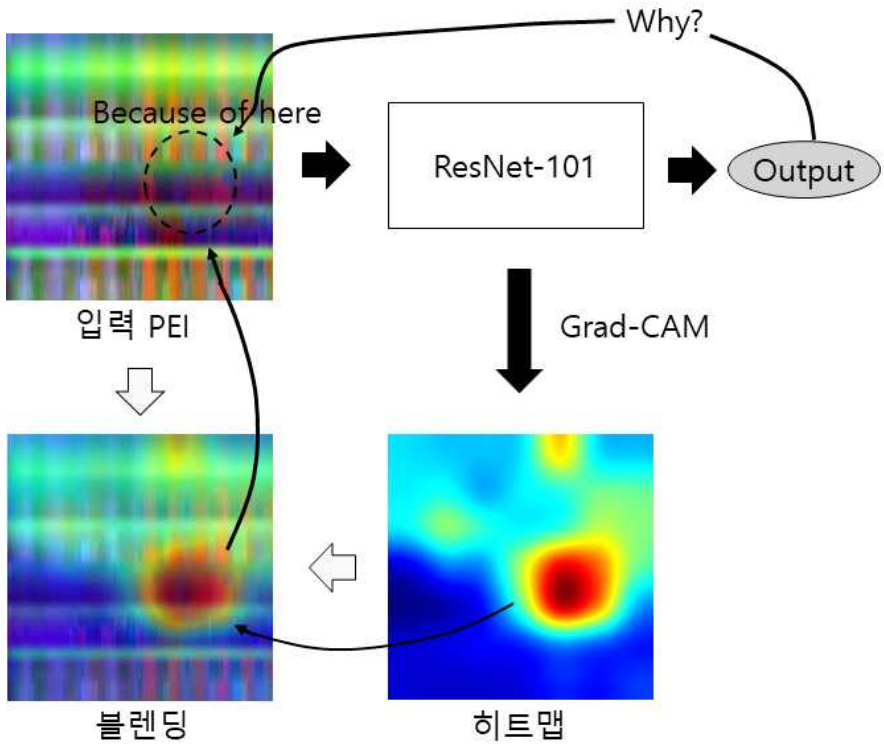


그림 3.32 Grad-CAM을 통한 히트맵과 의미

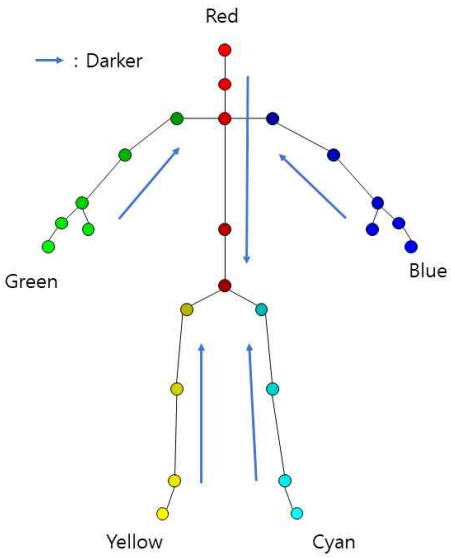


그림 3.33 히트궤적의 색상 배치

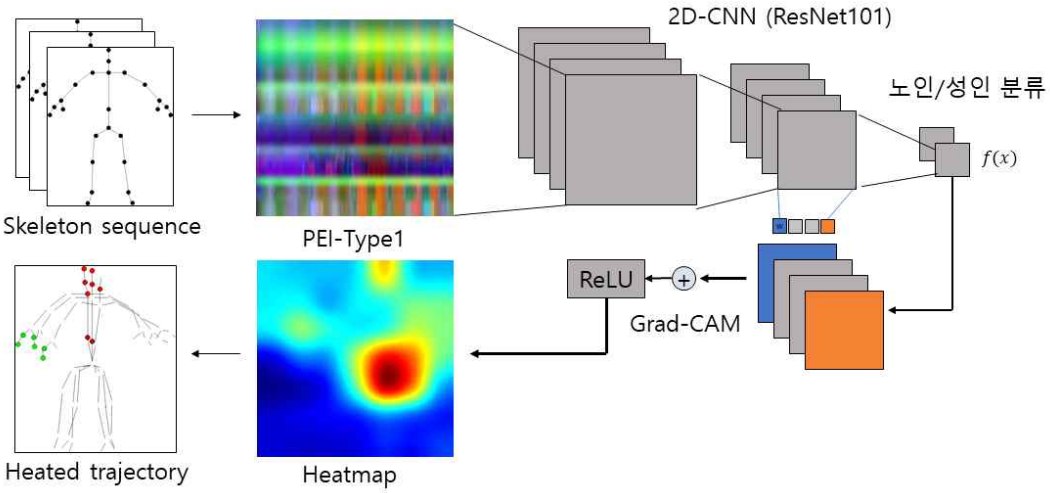


그림 3.34 행동특성분석의 방법

제4장 구현

실험에 앞서 본 장에서는 신경망의 학습 설정, 성능평가를 위한 평가 방법과 데이터셋을 기술한다. 실험에 사용한 장치의 구성으로 CPU(Central Processing Unit)는 Intel(R) Xeon(R) Gold 5120 2.2GHz, GPU(Graphics Processing Unit)는 NVIDIA Tesla V100-SXM2-32GB, RAM(Random Access Memory) 용량은 180GB, OS(Operating System)는 64비트기반 Window Server 2016를 사용하였다.

제1절 신경망의 학습 설정과 평가 척도

RGB (2D-CNN-LSTM) 모델의 구현에서 2D-CNN은 GoogLeNet과 ResNet-101을 고려하였고 ImageNet을 통해 학습된 초기 가중치 값들을 그대로 사용하여 마지막 풀링층의 특징벡터를 추출하였다. 한 비디오에서 프레임마다 특징벡터를 생성하고 이 특징벡터를 시퀀스로 만든다. 이상데이터를 학습에서 배제하기 위해 400길이의 시퀀스 데이터는 제외시켰다. LSTM은 2000개의 노드를 가지는 양방향 장단기 기억(BiLSTM) 계층과 과적합을 줄이는 드롭아웃(dropout)을 사용하여 설계하였고, 학습파라미터로서 최적화 방법은 Adam을 드롭아웃률은 0.5을 초기학습률은 0.0001을 학습 횟수(epoch)는 30을 미니배치 크기는 16을 학습률의 감쇠주기는 5를 학습률의 감쇠률은 0.2를 사용하여 학습하였다.

Skeleton (PEI-2D-CNN) 모델의 구현에서 2D-CNN은 ResNet-101을 고려하였고 ImageNet을 통해 학습된 초기 가중치 값들을 사용하여 전이학습을 하였다. 최적화 방법은 Adam, 미니배치크기는 30, 초기 학습률은 0.0001, 학습 횟수는 20을 학습률의 감쇠주기는 5를 학습률의 감쇠률은 0.2를 사용하여 학습하였다.

RGB (3D-CNN)은 특징추출 및 분류를 위한 3D-CNN을 R3D-18로 적용하였다. 학습 횟수는 50, 학습률은 0.001, 최적화방법은 Adam, 가중치감쇠(weight decay)는 0.00005, 미니배치 크기는 100을 사용하였다.

그 외 비교를 위한 방법들의 경우 오픈소스의 디폴트값을 그대로 사용하였으며 최적화방법은 Adam을 사용했다. 학습률은 가중치감쇠 전까지는 매 반복마다 1/3배부터 3배까지 무작위로 설정하였고 가중치감쇠부터는 0.001에서 0.000001까지 1/3 단위로 줄여나갔다. 미니배치 크기는 매 반복마다 GPU 메모리 최대치를 기준으로

1배에서 1/4배까지 무작위로 설정하여 다양하게 학습하였다[83].

표 4-1은 모델들의 검증 시간을 나타낸다. 384×216해상도의 327프레임의 비디오와 그 스켈레톤 시퀀스를 기준으로 측정하였다. 앙상블에 기여되는 개별 모델이 증가할수록 검증시간이 늘어나는 단점이 있다.

표 4-1. 모델들의 검증 시간

모델	검증시간 (fps)
RGB (2D-CNN-LSTM)	6.937s (47.1 f/s)
Skeleton (PEI-2D-CNN)	0.023s (14217.4 f/s)
RGB (3D-CNN)	3.571s (91.6 f/s)
Body ROI RGB (3D-CNN)	11.831s (27.6 f/s)
Hand-object ROI RGB (3D-CNN)	5.698s (57.4 f/s)
OpenPose	16.786s (19.5 f/s)
RGB-S-based 3-Stream Ensemble	10.531s (31.051 f/s)
ROI-based 4-Stream Ensemble	37.909s (8.626 f/s)

행동인식 모델의 성능평가 척도는 정확도(accuracy)로 옳은 분류수(N_c)를 옳은 분류수와 틀린 분류수(N_w)의 합으로 나눈 값이다. 식 4-1은 정확도를 계산하는 수식을 나타낸다.

$$accuracy = \frac{N_c}{N_c + N_w} \tag{4-1}$$

제2절 ETRI-Activty3D 데이터셋

본 논문에서 제안한 노인 행동특성 기반 심층 신경망을 이용한 행동인식의 성능을 평가하기 위해 ETRI(Electronics and Telecommunications Research Institute)-Activity3D 데이터셋을 사용하였다. 이 데이터는 총 112,620개 샘플로 구성되어 있는 두 번째로 큰 데이터셋이고, 노인 50명과 성인 50명으로부터 취득되었다. 노인은 64세부터 88세까지 평균 77.1세로 구성된 남자 17명과 여자 33명으로 구성되어 있고 성인은 21세부터 29세까지 평균 23.6세로 남자 25명과 여자 25명으로 구성되어 있다. 아파트 주거환경의 거실, 주방, 침실에서 일상생활 속 55가지 행동들을 수행하였고 키넥트 v2를 이용해 취득되었다. 여기서 55가지 행동은 노인들의 일상생활에서 자주 일어나는 행동들을 관찰하여 정의한 것들이다. 가정용 서비스 상황을 가정하여 키넥트 센서는 70cm와 120cm 높이에서 4대씩 구성하여 8가지 방향에서 획득하였다. 촬영장치와 대상자의 거리는 1.5m에서 3.5m 이내로 취득하였다. 취득된 데이터의 형식은 컬러이미지의 경우 1920×1080, 뎁스이미지의 경우 512×424 해상도이고 스켈레톤 정보는 3차원 공간의 25가지 관절 위치를 포함한다. 데이터의 프레임레이트는 20이다. 표 4-2은 ETRI-Activity 3D 데이터의 행동 종류를 보여주고 그림 4.1은 ETRI-Activity3D의 데이터 예시를 나타낸다. 데이터의 다양성을 위해서 한 행동에서 한 사람이 집안의 장소(거실, 침실, 주방 등) 또는 정면방향을 바꾸며 2~3번 행동한 것을 공간조건에 따라 동시에 4대 또는 8대에서 100명에 대해 획득한다. 행동마다 평균 약 2050개 데이터가 있고 그 안에서 한 명당 평균 20.5개 데이터를 가진다[83]. 데이터의 전체 크기가 너무 크기 때문에 해상도를 1/5로 낮추어 384×216해상도를 가지도록 다운사이즈하였다.

표 4-2. ETRI-Activity3D의 행동 종류

1	포크로 음식 먹기	29	세탁물 널기
2	컵에 물 붓기	30	무언가 찾기
3	약 먹기	31	리모컨 사용하기
4	물 마시기	32	책 읽기
5	냉장고에서 음식 꺼내거나 넣기	33	신문 읽기
6	채소 다듬기	34	글쓰기
7	과일 깎기	35	전화기에 말하기
8	가스 불 켜기	36	휴대전화로 놀기
9	도마에서 채소 썰기	37	컴퓨터 하기
10	이 닦기	38	담배 피우기
11	손 씻기	39	박수 치기
12	세수하기	40	손으로 얼굴 만지기
13	수건으로 얼굴 닦기	41	맨손체조하기
14	화장하기	42	고개 돌리기
15	립스틱 바르기	43	혼자 어깨 주무르기
16	머리카락 빗기	44	인사하기
17	머리카락 드라이하기	45	서로 대화하기
18	재킷 입기	46	악수하기
19	재킷 벗기	47	서로 포옹하기
20	신발 신거나 벗기	48	서로 싸우기
21	안경 쓰거나 벗기	49	손 흔들기
22	설거지하기	50	오라고 손짓하기
23	바닥 진공 청소하기	51	손가락으로 가리키기
24	걸레로 바닥 닦기	52	문 열고 들어가기
25	식탁 닦기	53	바닥에 쓰러지기
26	가구 닦기	54	앉거나 일어서기
27	침구 접거나 펴기	55	눕기
28	손으로 수건 빨기		

그림 4.1 ETRI-Activity3D 데이터 예시

제3절 평가 방법

기존 연구 규격에 따라 데이터셋을 CS(Cross-Subject)와 CA(Cross-Age)로 나누어 평가한다. ETRI-Activity3D를 기준으로, CS는 노인 50명을 1부터 50까지, 성인 50명을 51부터 100까지라고 할 때 1부터 3배수를 제외한 번호는 학습 데이터로, 1부터 3배수는 검증 데이터로 분리하였다. 학습 데이터에 성인과 노인이 혼합되어 67명, 검증 데이터에 성인과 노인이 혼합되어 33명으로 구성된다. CA는 노인과 성인을 분리하여 구성한 것으로 노인학습, 노인검증, 성인학습, 성인검증으로 구성된다. CA는 CS와 동일하게 노인 50명을 1부터 50까지, 성인을 51부터 100까지라고 할 때 1부터 3배수를 제외한 번호는 학습 데이터로, 1부터 3배수는 검증 데이터로 분리하되 50번과 51번사이의 노인과 성인의 경계선에서 도메인을 분리한다[83]. 또한, CS의 교차검증(cross validation)으로 노인 50명을 1부터 50까지, 성인 50명을 51부터 100까지라고 할 때 2부터 3배수를 제외한 번호는 학습 데이터로, 2부터 3배수는 검증 데이터로 분리하였다. 데이터가 크기 때문에 1번의 교차검증을 수행하였다. 그림 4.2는 ETRI-Activity3의 Cross-Subject셋의 구성을 보여주고 그림 4.3은 ETRI-Activity3의 Cross-Age셋의 구성을 보여준다. 그림 4.4는 교차검증을 위한 ETRI-Activity3의 Cross-Subject셋의 구성을 보여준다.

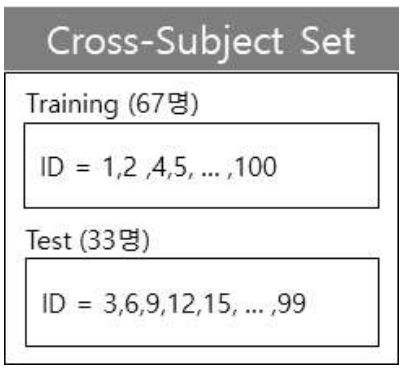


그림 4.2 ETRI-Activity3D의 CS(Cross-Subject)셋의 구성

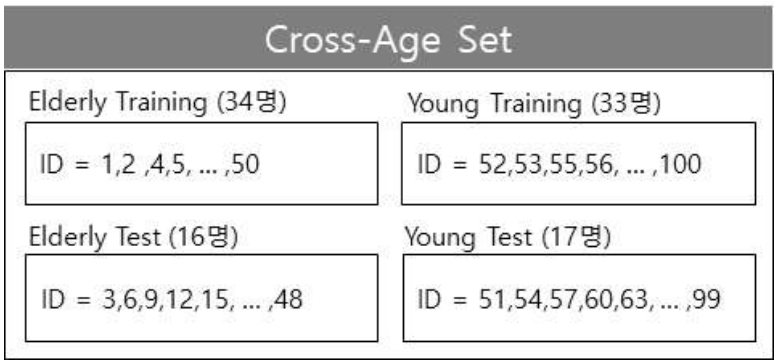


그림 4.3 ETRI-Activity3D의 CA(Cross-Age)셋의 구성

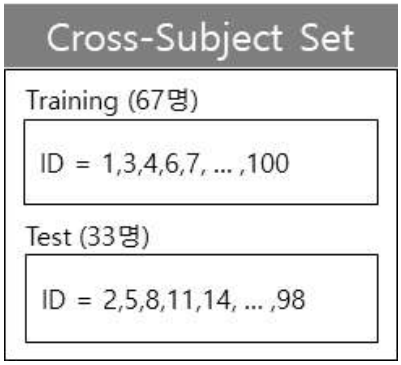


그림 4.4 교차검증을 위한 ETRI-Activity3의 Cross-Subject셋의 구성

5장 실험 및 결과분석

본 장에서는 제안한 방법으로 ETRI-Activity3D의 55가지 행동을 분류하여 기존 방법보다 정확도가 향상되는지 확인하고 제안한 행동특성분석 방법으로 행동특성 분석이 가능한지 확인한다.

제1절 비디오와 스켈레톤의 앙상블기반 행동인식 실험

행동인식 데이터셋 동향에서 데이터들이 보통 RGB영상과 깊이영상, 스켈레톤 등으로 구성되어 있다. 하나의 디바이스 안에 이러한 데이터를 취득하는 센서들이 내장되어 동시에 기록이 가능하다. 다양한 센서정보는 데이터를 분석하는데 더 많은 정보를 주기 때문에 정확도를 향상시키는데 도움을 줄 수 있다. 그러한 이유로 RGB비디오 영상과 스켈레톤 시퀀스를 개별적으로 사용한 인식기들을 설계하고 마지막에 각 인식기들의 스코어들을 앙상블하여 각 데이터들의 특성을 최적으로 조화시킬 수 있다. 우선 RGB비디오를 이용한 인식기와 스켈레톤 시퀀스를 이용한 인식기로 구분한다. RGB비디오는 2D-CNN-LSTM 방법과 3D-CNN을 사용하여 인식기를 설계하고, 스켈레톤 시퀀스는 PEI-2D-CNN방법을 사용하여 인식기를 설계한다.

2D-CNN-LSTM 방법은 RGB비디오의 매 프레임을 2D-CNN에 입력하여 특징벡터들을 구하고 그 특징벡터들을 다시 시퀀스로 구성하여 LSTM에 입력하여 분류한다. 이때 비디오마다 영상의 길이가 달라서 일반적으로 정규화를 추가로 해주어야 하는데 LSTM은 다양한 입력 길이를 받아들일 수 있을 뿐만 아니라 입력된 순서에 포함된 특징을 잘 추출할 수 있다. 2차원 영상특징추출기로서 2D-CNN은 사전학습 모델인 GoogLeNet과 ResNet101을 적용하였고 초기 파라미터 그대로 사용하였다. 사전학습 모델은 이미 좋은 구조임을 검증받고 학습된 공개된 모델로서 신경망 특성상 다양한 성분과 깊이의 구조로 설계할 수 있고 그 구조에 따라 성능도 제각각이라 시행착오를 많이 겪어야 하는 과정을 줄이고 빠르게 적용할 수 있다. 또한, 수십에서 수백에 이르는 레이어들을 일일이 정의해야 하는 번거로움을 없앨 수 있다. LSTM은 2000개의 노드를 가지는 양방향 장단기 기억(BiLSTM) 계층과 과적합을 줄이는 드롭아웃(dropout)을 사용하여 설계하였고, 학습파라미터로서 최적화 방법은 Adam을 초기학습률을 0.0001을 학습횟수는 30을 미니배치 크기는 16을 적용하였다.

3D-CNN 방법은 입력 데이터의 차원이 3차원인 CNN으로 특징추출을 위해 내부에 설계된 필터들 역시도 3차원 형태를 가진다. 3차원 합성곱 연산은 공간영역뿐만 아니라 시간영역에 특징도 같이 계산되기 때문에 시퀀스 데이터에 2D-CNN보다 더 잘 학습할 수 있다. 3차원 영상특징추출기 및 분류기로서 3D-CNN은 사전학습 모델인 R3D-18을 사용하였다. 학습횟수는 50, 학습률은 0.001, 최적화방법은 Adam, 가중치감쇠(weight decay)는 0.00005, 미니배치 크기는 100을 사용하였다. 그림 5.1은 RGB비디오 데이터의 예시를 보여주고 표 5-1는 RGB기반 행동인식의 정확도(CS)를 보여준다. RGB (2D-CNN-LSTM-Type1)은 특징추출을 위한 2D-CNN을 GoogLeNet, RGB (2D-CNN-LSTM-Type2)은 특징추출을 위한 2D-CNN을 ResNet-101, RGB (3D-CNN)은 특징추출 및 분류를 위한 3D-CNN을 R3D-18로 적용하였다. 그리고 그림 5.2과 그림 5.3는 RGB비디오 입력의 2D-CNN-LSTM-Type1 학습과정과 RGB비디오 입력의 2D-CNN-LSTM-Type2 학습과정을 보여주고 그림 5.4은 RGB비디오 입력의 3D-CNN 학습과정을 나타낸다.

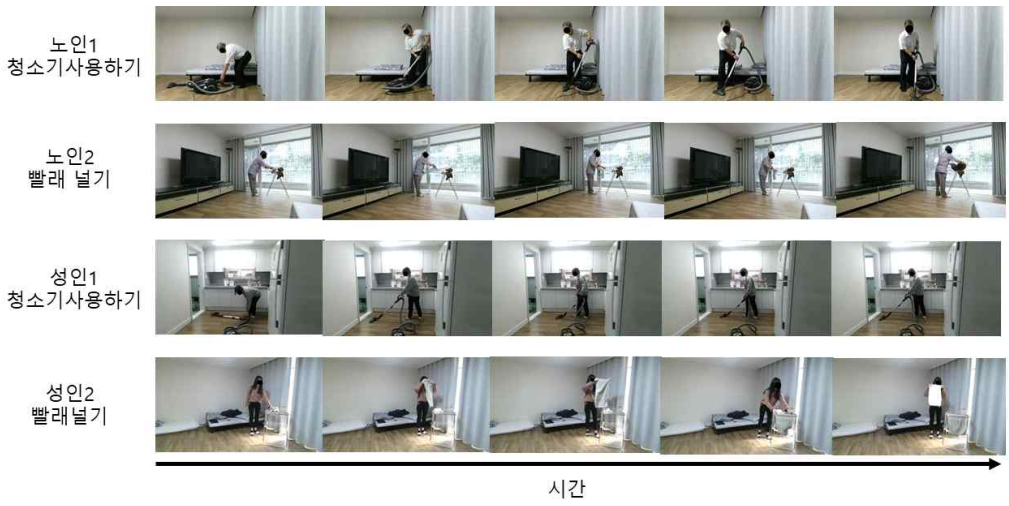
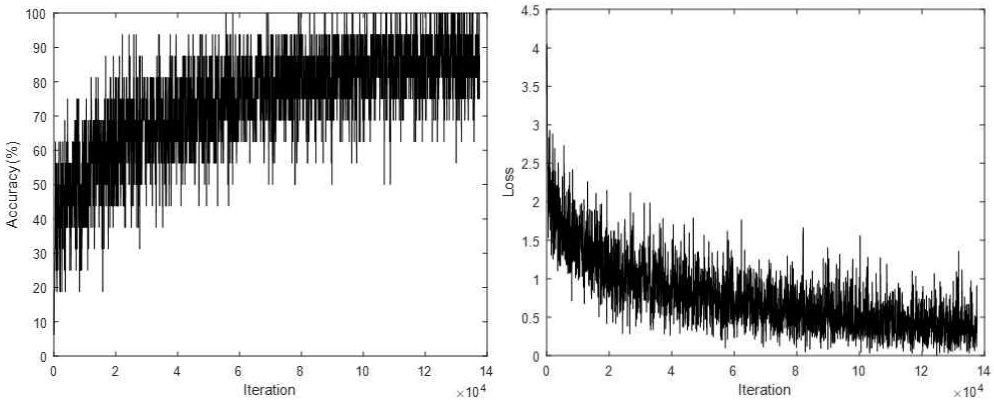


그림 5.1 RGB비디오 데이터의 예시

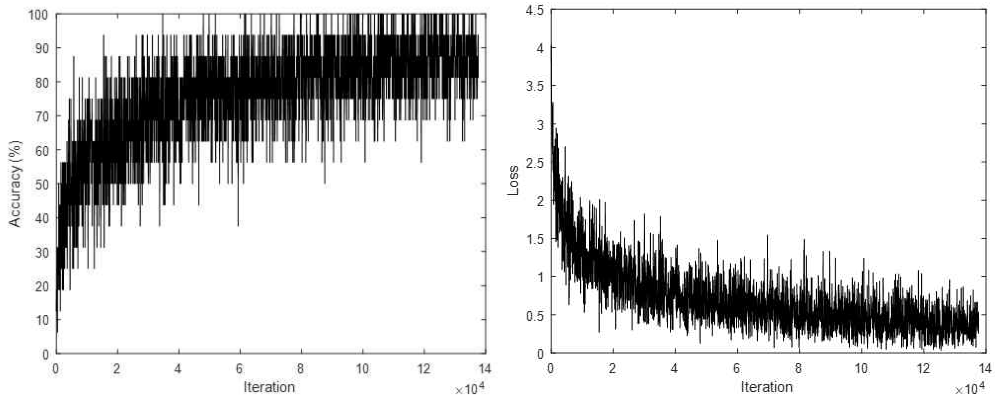
표 5-1. RGB기반 행동인식의 정확도(CS)

방 법		정확도(%)
RGB-based network	2D-CNN-LSTM-Type1	49.47
	2D-CNN-LSTM-Type2	47.49
	3D-CNN	79.20



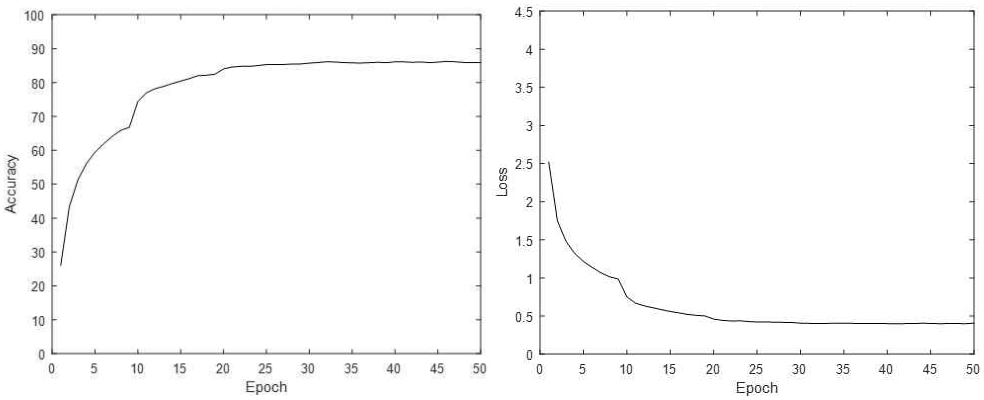
(a) 정확도 (b) 손실

그림 5.2 RGB비디오 입력의 2D-CNN-LSTM-Type1 학습과정



(a) 정확도 (b) 손실

그림 5.3 RGB비디오 입력의 2D-CNN-LSTM-Type2 학습과정



(a) 정확도 (b) 손실

그림 5.4 RGB비디오 입력의 3D-CNN 학습과정

PEI-T1-2D-CNN 방법은 3차원의 스켈레톤 시퀀스를 PEI 방법을 통해 이미지로 변환한 후 그 이미지를 2D-CNN에 입력하여 분류하는 것으로 PEI 타입1~4의 유형을 T1~4로 표기하였다. 변환 된 이미지는 224×224×3 크기의 RGB이다. 2D-CNN은 사전학습 모델인 ResNet101을 사용하였다. 최적화 방법은 Adam, 미니배치크기는 30, 초기 학습률은 0.0001, 학습 횟수(epoch)는 20을 사용하였다. 그림 5.5부터 그림 5.8는 PEI-T1, PEI-T2, PEI-T3, PEI-T4의 예시를 보여주고 표 5-2는 스켈레톤기반 행동인식 정확도(CS)를 보여준다. 그림 5.9부터 그림 5.12은 PEI-T1-2D-CNN, PEI-T2-2D-CNN, PEI-T3-2D-CNN, PEI-T4-2D-CNN의 학습과정을 나타낸다.

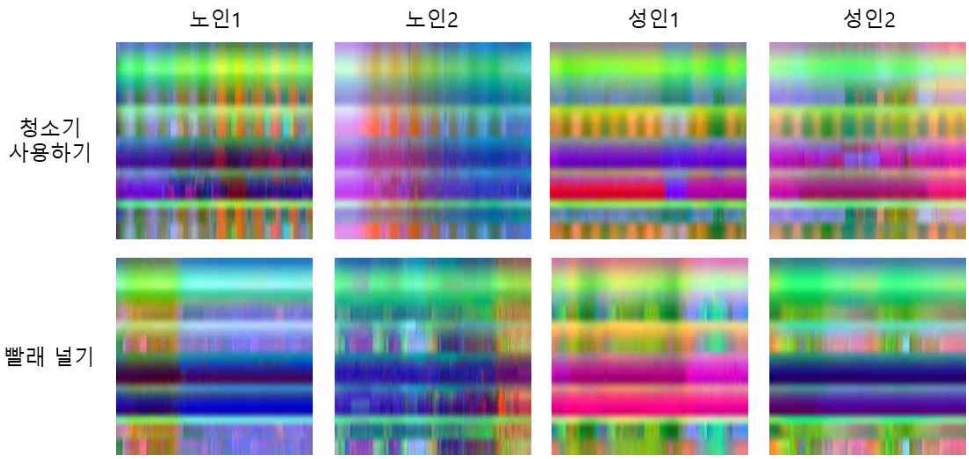


그림 5.5 PEI-T1의 예시

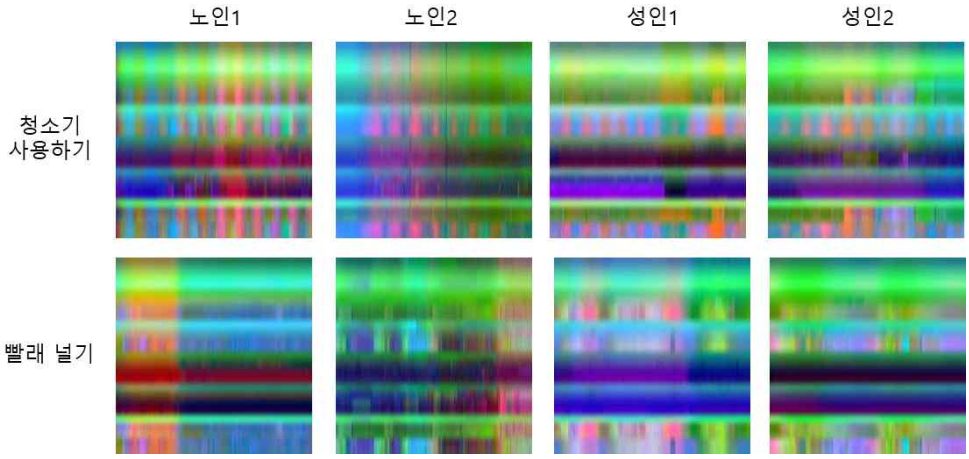


그림 5.6 PEI-T2의 예시

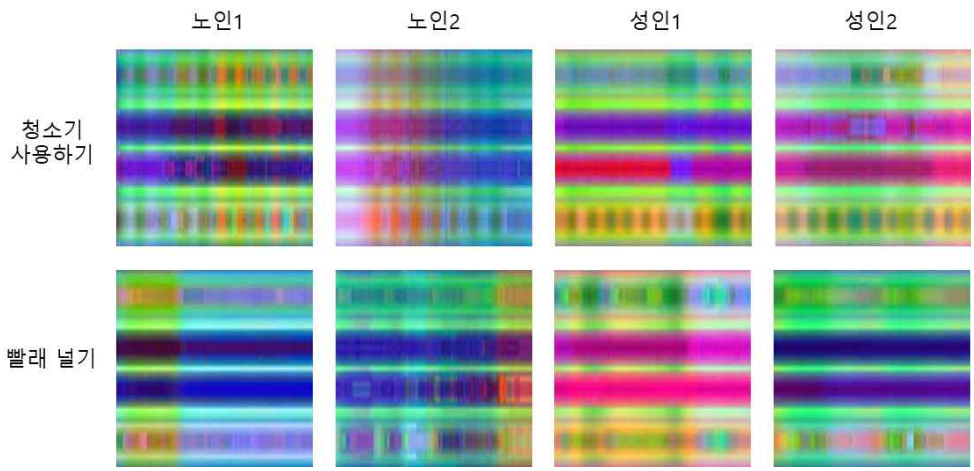


그림 5.7 PEI-T3의 예시

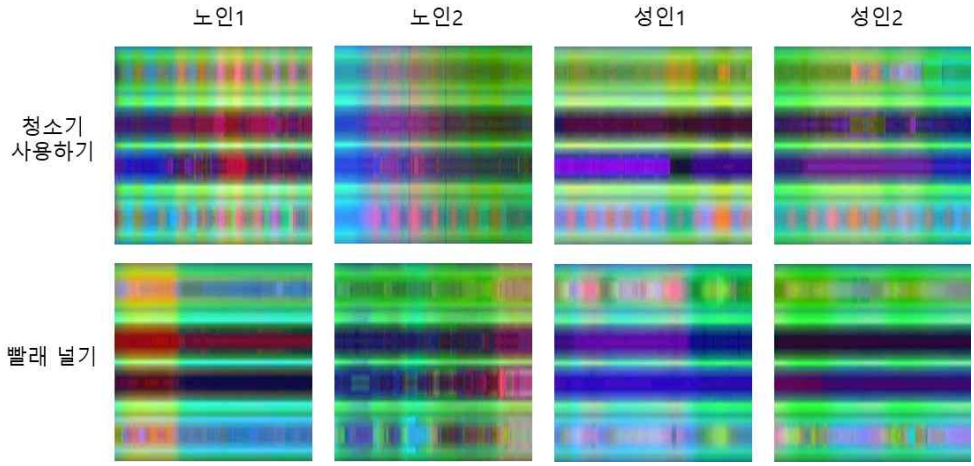
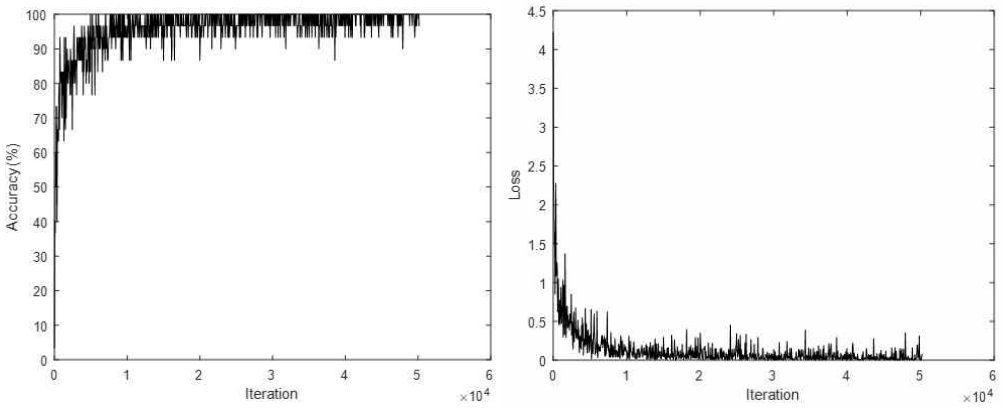


그림 5.8 PEI-T4의 예시

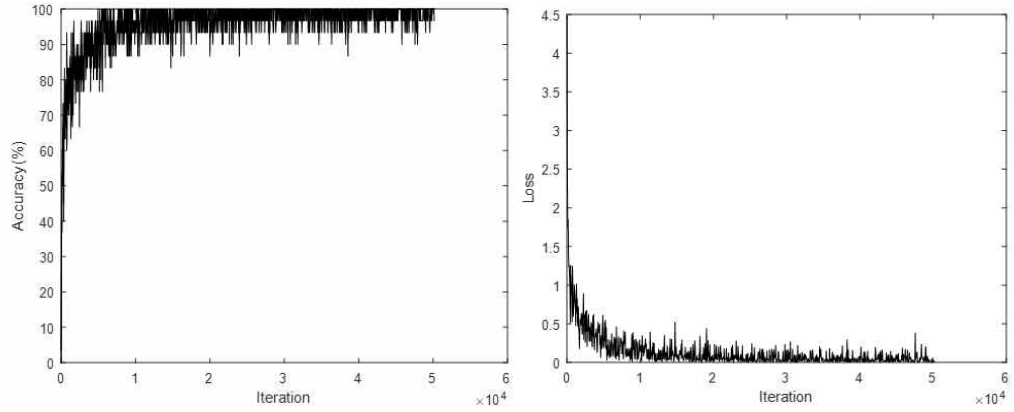
표 5-2. 스켈레톤기반 행동인식 정확도(CS)

방 법		정확도(%)
Skeleton-based network	PEI-T1-2D-CNN	84.95
	PEI-T2-2D-CNN	85.88
	PEI-T3-2D-CNN	86.09
	PEI-T4-2D-CNN	85.20



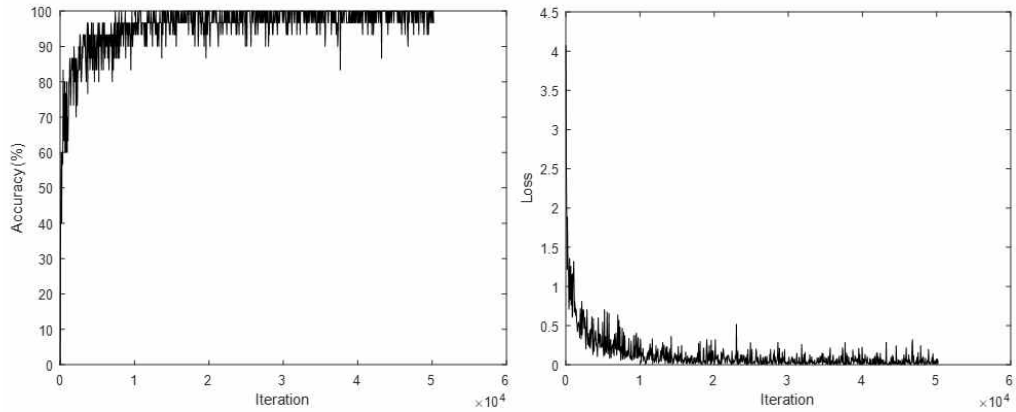
(a) 정확도 (b) 손실

그림 5.9 PEI-T1-2D-CNN 학습과정



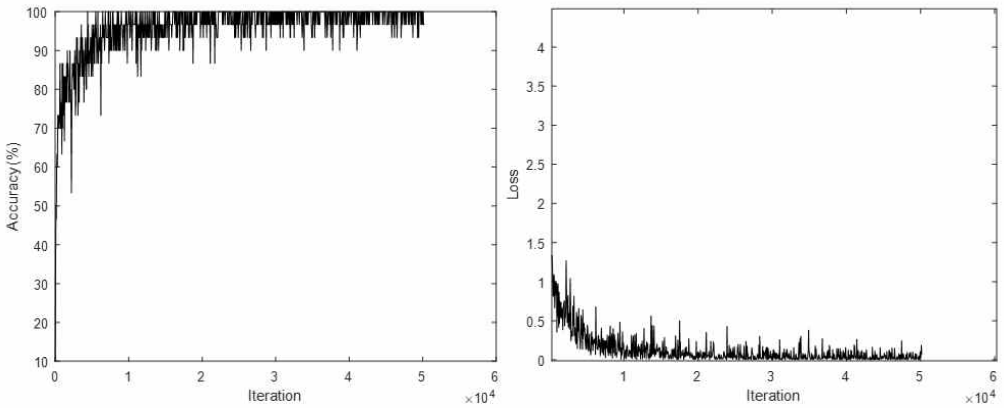
(a) 정확도 (b) 손실

그림 5.10 PEI-T2-2D-CNN 학습과정



(a) 정확도 (b) 손실

그림 5.11 PEI-T3-2D-CNN 학습과정



(a) 정확도 (b) 손실

그림 5.12 PEI-T4-2D-CNN 학습과정

앞서 스켈레톤기반 행동인식은 PEI를 생성할 때 스켈레톤의 방향과 관절의 수에 변화를 주어 정보를 다양하게 만들었기 때문에 이들의 분류 결과를 앙상블하여 분류성능을 높일 수 있다. 앙상블 방법으로는 여러 분류기 모델의 출력 스코어의 값들을 더하거나 곱해서 최대치로 최종 분류하였다. 그림 5.13는 스켈레톤기반 행동인식 모델들의 앙상블을 보여주고 표 5-3는 스켈레톤기반 행동인식의 앙상블 정확도(CS)를 보여준다. Skeleton-based ensemble network (Type1)는 스켈레톤 데이터의 PEI영상을 입력으로 PEI-T1-2D-CNN, PEI-T2-2D-CNN, PEI-T3-2D-CNN, PEI-T4-2D-CNN의 각 출력 스코어 값들을 더해서 앙상블 한 것이고, 마찬가지로 Skeleton-based ensemble network (Type2)는 스켈레톤 데이터의 PEI영상을 입력으로 PEI-T1-2D-CNN, PEI-T2-2D-CNN, PEI-T3-2D-CNN, PEI-T4-2D-CNN의 각 출력 스코어 값들을 곱해서 앙상블 한 것이다.

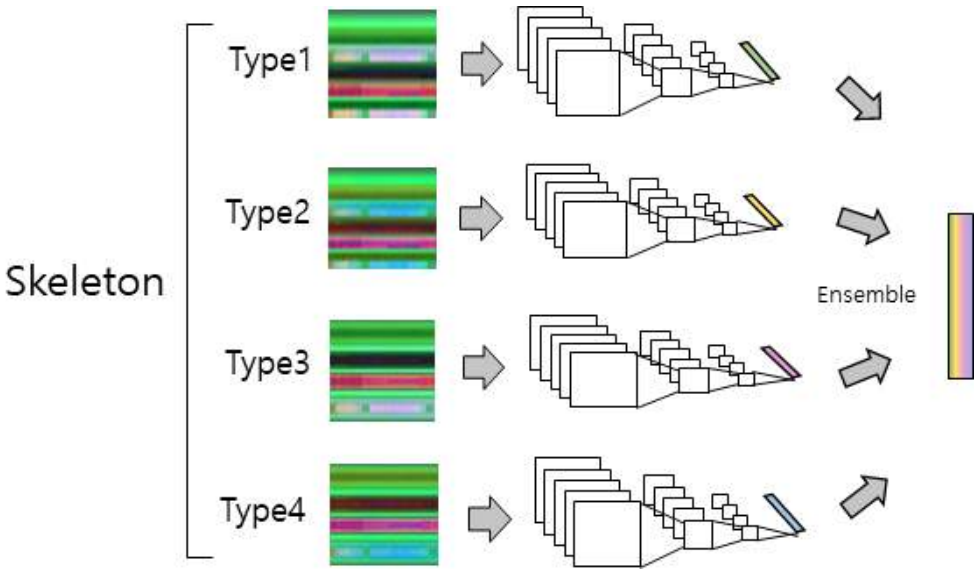


그림 5.13 스켈레톤기반 행동인식 모델들의 앙상블

표 5-3. 스켈레톤기반 행동인식의 앙상블 정확도(CS)

방 법		정확도(%)
Skeleton-based ensemble network	Type1	90.13
	Type2	90.42

앞서 RBG기반 행동인식과 스켈레톤기반 행동인식은 서로 다른 유형의 데이터로 각각 가지고 있는 정보가 다르고 분류기 모델 또한 중점적으로 추출하는 특징이 다르기 때문에 이들의 분류 결과를 앙상블하여 분류성능을 극대화 할 수 있다. 앙상블 방법으로는 여러 분류기 모델의 출력 스코어의 값들을 더하거나 곱해서 최대치로 최종 분류하였다. 표 5-4는 RGB와 스켈레톤기반 행동인식의 앙상블 정확도(CS)를 보여준다. RBG-S기반 3-스트림 앙상블 모델로서 RGB-S-based ensemble network (Type1)는 RGB영상을 입력으로 3D-CNN, RGB영상을 입력으로 2D-CNN-LSTM, 스켈레톤 데이터의 PEI-T3 영상을 입력으로 2D-CNN의 출력 스코어 값들을 더해서 앙상블 한 것이고 마찬가지로 RGB-S-based ensenble network (Type2)는 RGB영상을 입력으로 3D-CNN, RGB영상을 입력으로 2D-CNN-LSTM, 스켈레톤 데이터의 PEI-T3 영

상을 입력으로 2D-CNN의 출력 스코어 값들을 곱해서 앙상블 한 것이다. 3D-CNN으로는 R3D-18, LSTM의 2D-CNN으로는 GoogLeNet, PEI-T3의 2D-CNN으로는 ResNet-101을 사용하였다. RGB와 스켈레톤 데이터를 앙상블 했을 때 단독 모델보다 최소 7.11%에서 최대 43.73%까지 향상된 것을 확인할 수 있다.

표 5-4. RGB와 스켈레톤기반 행동인식의 앙상블 정확도(CS)

방 법		정확도(%)
RGB-S-based ensemble network	Type1	91.02
	Type2	93.20

RGB-S기반 3-스트림 앙상블 모델의 성능 안정성을 평가하기 위해 앞서 CS에서 학습에서 사용한 데이터를 검증으로, 검증에서 사용한 데이터를 학습으로 재구성하여 CS-교차검증을 수행하였다. 데이터 규모가 거대하기 때문에 단일의 CS-교차검증을 수행하였다. 학습 옵션 및 구현 환경은 앞서 CS의 조건과 똑같다. 표 5-5는 RGB 또는 스켈레톤기반 행동인식의 정확도(CS-교차검증)을 보여주고, 표 5-6은 RGB와 스켈레톤기반 행동인식의 앙상블 정확도(CS-교차검증)을 나타낸다.

표 5-5. RGB 또는 스켈레톤기반 행동인식의 정확도(CS-교차검증)

방 법		정확도(%)
Skeleton-based network	PEI-T3-2D-CNN	85.17
RGB-based network	2D-CNN-LSTM-Type1	48.80
	3D-CNN	77.85

표 5-6. RGB와 스켈레톤기반 행동인식의 앙상블 정확도(CS-교차검증)

방 법		정확도(%)
RGB-S-based ensemble network	Type1	90.36
	Type2	92.76

RGB-S기반 3-스트림 앙상블 모델의 성능 안정성 평가 및 도메인 차이 분석을 위해 앞서 CS를 노인과 성인으로 분할시켜 CA 검증을 수행하였다. 학습 옵션 및 구현 환경은 앞서 CS의 조건과 똑같다. 표 5-7는 RGB 또는 스켈레톤기반 행동인식의 정확도(CA)을 보여주고, 표 5-8은 RGB와 스켈레톤기반 행동인식의 앙상블 정확도(CA)을 나타낸다. Case1은 노인학습인 경우, Case2는 성인학습인 경우를 의미한다.

표 5-7. RGB 또는 스켈레톤기반 행동인식의 정확도(CA)

방 법			정확도(%)	
			노인	성인
Skeleton-based network	PEI-T3-2D-CNN	Case1	85.11	66.80
		Case2	70.37	83.27
RGB-based network	2D-CNN-LSTM-Type1	Case1	49.72	29.06
		Case2	28.18	42.73
	3D-CNN	Case1	77.25	55.18
		Case2	47.16	83.47

표 5-8. RGB와 스켈레톤기반 행동인식의 앙상블 정확도(CA)

방 법			정확도(%)	
			노인	성인
RGB-S-based ensemble network	Type1	Case1	90.78	68.96
		Case2	70.96	87.16
	Type2	Case1	92.81	73.56
		Case2	75.08	90.37

제2절 Body ROI와 Hand-object ROI기반 행동인식 실험

ROI(Region of Interest)기반 행동인식은 RGB비디오를 분석하기 전에 이미지 전체 영역 중에 관심있는 영역만 남겨두고 인식을 수행하는 것이다. Body ROI는 이미지 전체에서 사람만 남겨두어 관심 대상에 초점을 맞추는 방법이고 또 다른 Hand-object ROI는 이미지 전체에서 사람이 사용하는 도구에 초점을 맞추기 위해 손 영역만 남겨두는 방법이다. 사람부분만 남겨두기 위해서 우선 오픈포즈를 이용해 2차원 관절 좌표들을 얻고 그 좌표를 중심으로 일정크기의 폭만큼 잘라내어 배경이 없는 빈 동일 크기의 이미지의 같은 좌표에 대입한다. 모든 관절에 대해 이 과정을 거치게 되면 사람부분만 남고 나머지는 검은색으로 나타난다. 손의 물체의 경우는 앞서처럼 모든 관절대신 손 관절에 대해서만 수행한다. 모든 RGB비디오에 대해 이 처리를 하게 되면 ROI기반 행동인식을 할 준비가 완료된다. 준비된 데이터를 3D-CNN에 입력으로 넣은 후 학습시켜 인식한다. 3차원 데이터의 특징추출기 및 분류기로서 3D-CNN은 사전학습 모델인 R3D-18을 사용하였다. 학습횟수는 50, 학습률은 0.001, 최적화방법은 Adam, 가중치감쇠는 0.00005, 미니배치 크기는 100을 사용하였다. 행동인식 모델의 성능평가 방법은 정확도를 사용한다. 그림 5.14은 Body ROI 비디오 데이터의 예시를 보여주고 그림 5.15은 Hand-object ROI 비디오 데이터의 예시를 보여준다. 표 5-9는 ROI기반 행동인식의 정확도(CS)를 보여주고 그림 5.16은 Body ROI기반 3D-CNN 학습과정을 나타낸다. 그림 5.17은 Hand-object ROI기반 3D-CNN 학습과정을 나타낸다.

그림 5.14 Body ROI 비디오 데이터의 예시

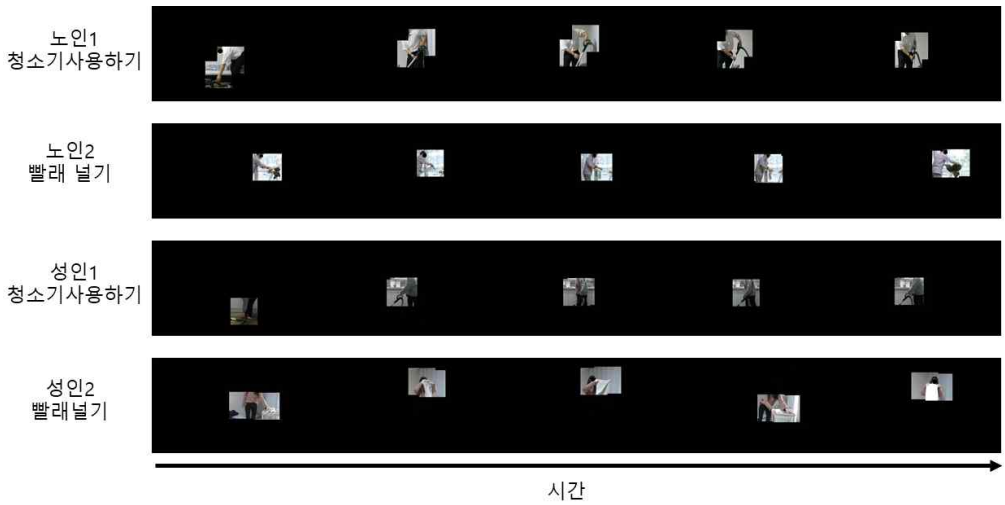
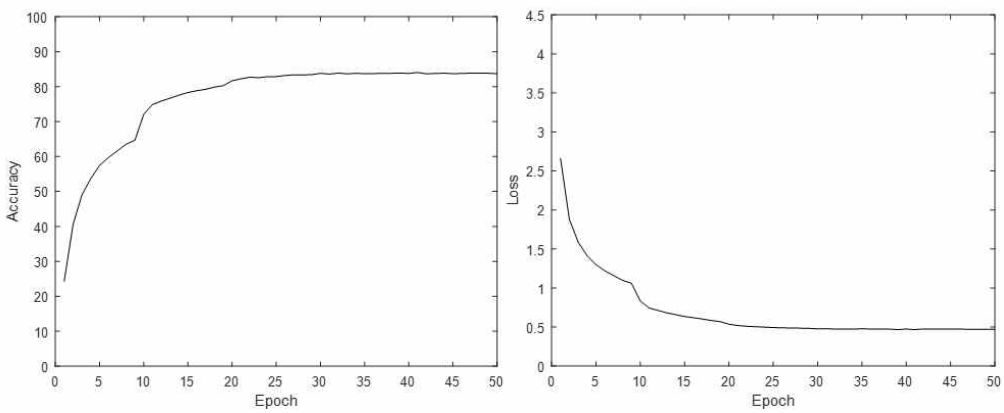


그림 5.15 Hand-object ROI 비디오 데이터의 예시

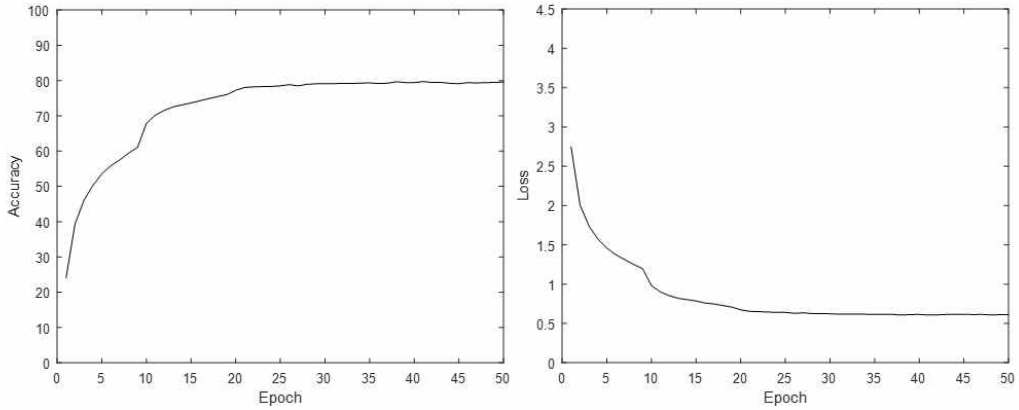
표 5-9. ROI기반 행동인식 정확도(CS)

방 법		정확도(%)
RGB-based network	3D-CNN	79.20
Hand-object ROI-based network		73.11
Body ROI-based network		76.85



(a) 정확도 (b) 손실

그림 5.16 Body ROI 기반 3D-CNN 학습과정



(a) 정확도 (b) 손실

그림 5.17 Hand-object ROI기반 3D-CNN 학습과정

RBG비디오에서 행동인식에 주요 정보를 가지는 Body ROI에 집중하여 3D-CNN을 새롭게 학습하므로 이미지의 전체부분으로 학습시킬 때와는 또 다른 특징들이 추출되고 다른 분석을 하게 된다. 또한 RGB비디오에서 행동인식에 주요 정보를 가지는 Hand-object ROI에 집중하여 새롭게 학습하므로 이미지의 전체부분으로 학습시킬 때와는 또 다른 특징들이 추출되고 다른 분석을 하게 된다. ROI는 RGB에서 정보를 지웠기 때문에 단독적으로 성능이 다소 낮아질 수 있지만 이러한 다양한 분석의 결과의 정보를 앙상블하면 단일 성능보다 훨씬 뛰어넘는 성능을 기대할 수 있다. 표 5-10은 ROI기반 모델의 앙상블 결과(CS)를 보여준다. ROI-based ensemble network의 Type1과 Type2는 RGB비디오 입력의 3D-CNN과 Body ROI 입력의 3D-CNN의 결과들을 각각 덧셈과 곱셈 앙상블한 모델이고 ROI-based ensemble network의 Type3과 Type4는 RGB비디오 입력의 3D-CNN, Body ROI 입력의 3D-CNN, 그리고 Hand-object ROI 입력의 3D-CNN의 결과들을 각각 덧셈과 곱셈 앙상블한 모델이다. ROI-based ensemble network의 Type5와 Type6은 RGB비디오 입력의 3D-CNN, Body ROI 입력의 3D-CNN, Hand-object ROI 입력의 3D-CNN, PEI-T3-2D-CNN의 결과들을 각각 덧셈과 곱셈 앙상블한 모델이고 ROI-based ensemble network의 Type7와 Type8은 RGB비디오 입력의 3D-CNN, Body ROI 입력의 3D-CNN, Hand-object ROI 입력의 3D-CNN, PEI-T1~T4-2D-CNN의 결과들을 각각 덧셈과 곱셈 앙상블한 모델이다. 실험결과 ROI-based ensemble network (Type6)이 다른 조합의 앙상블 결과들보다 가장 높은 정확도를 가졌다.

표 5-10. ROI기반 모델의 앙상블 결과(CS)

방 법		정확도(%)
ROI-based ensemble network	Type1	84.68
	Type2	85.85
	Type3	86.83
	Type4	87.98
	Type5	92.79
	Type6	94.87
	Type7	94.18
	Type8	94.69

그림 5.18과 표 5-11은 기존 행동인식 방법들과 제안한 행동인식 방법을 비교한 결과(CS)를 보여준다. 제안한 RGB-S-based ensemble network (Type2)와 ROI-based ensemble network (Type6)이 다른 기존 방법보다 최소 2.6%에서 최대 20.97% 더 성능 개선된 것을 확인 할 수 있다.

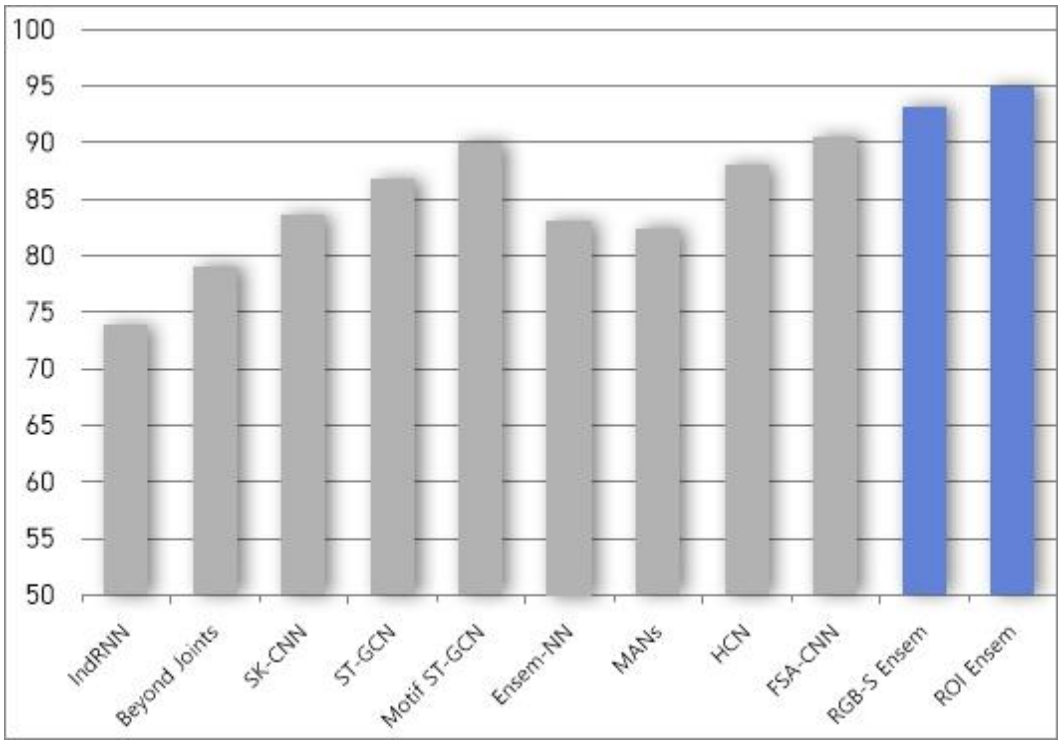


그림 5.18 기존 행동인식 방법들과 성능비교(CS)

표 5-11. 기존 행동인식 방법들과 성능비교(CS)

방법	ETRI-Activity3D
IndRNN[20]	73.90%
Beyond Joints[21]	79.10%
SK-CNN[22]	83.60%
ST-GCN[23]	86.80%
Motif ST-GCN[24]	89.90%
Ensem-NN[25]	83.00%
MANs[26]	82.40%
HCN[27]	88.00%
FSA-CNN[83]	90.60%
제안된 RGB-S-based ensemble network (Type2)	93.20%
제안된 ROI-based ensemble network (Type6)	94.87%

ROI기반 4-스트림 앙상블 모델의 성능 안정성을 평가하기 위해 앞서 CS에서 학습에서 사용한 데이터를 검증으로, 검증에서 사용한 데이터를 학습으로 재구성하여 CS-교차검증을 수행하였다. 데이터 규모가 거대하기 때문에 단일의 CS-교차검증을 수행하였다. 학습 옵션 및 구현 환경은 앞서 CS의 조건과 똑같다. 표 5-12는 ROI기반 행동인식의 정확도(CS-교차검증)를 보여주고, 표 5-13은 ROI기반 모델의 앙상블 결과(CS-교차검증)를 나타낸다.

표 5-12. ROI기반 행동인식의 정확도(CS-교차검증)

방 법		정확도(%)
Hand-object ROI-based network	3D-CNN	73.22
Body ROI-based network		76.12

표 5-13. ROI기반 모델의 앙상블 결과(CS-교차검증)

방 법		정확도(%)
ROI-based ensemble network	Type5	92.35
	Type6	94.23

ROI기반 4-스트림 앙상블 모델의 성능 안정성 평가 및 도메인 차이 분석을 위해 앞서 CS를 노인과 성인으로 분할시켜 CA 검증을 수행하였다. 학습 옵션 및 구현 환경은 앞서 CS의 조건과 똑같다. 표 5-14는 ROI기반 행동인식의 정확도(CA)을 보여주고, 표 5-15은 ROI기반 모델의 앙상블 결과(CA)을 나타낸다. Case1은 노인 학습인 경우, Case2는 성인학습인 경우를 의미한다.

표 5-14. ROI기반 행동인식의 정확도(CA)

방 법			정확도(%)	
			노인	성인
Hand-object ROI-based network	3D-CNN	Case1	71.04	52.55
		Case2	43.57	77.40
Body ROI-based network	3D-CNN	Case1	74.56	53.30
		Case2	48.35	80.43

표 5-15. ROI기반 모델의 앙상블 결과(CA)

방 법			정확도(%)	
			노인	성인
ROI-based ensemble network	Type5	Case1	92.53	70.35
		Case2	73.57	89.87
	Type6	Case1	94.57	75.04
		Case2	79.51	92.54

제3절 설명 가능한 AI를 이용한 행동특성분석 실험

사람은 나이가 들수록 신체적 구조가 바뀌고 신체적 기능이 떨어지므로 성인과 노인간에 행동특성에 변화가 생긴다. 노인은 시각, 청각 및 근력이 떨어져 외부 환경을 지각하는데 더디고 반응이 성인에 비해 느린 특성을 보인다. 노인과 성인의 차이를 시각화하기 위해 t-SNE(Stochastic Neighbor Embedding)를 적용한다. 특정 단일 행동에 대해서 스켈레톤 시퀀스를 PEI-type1으로 변환하고 ResNet-101을 노인/성인을 분류하도록 학습시킨다. 최적화 방법은 Adam, 미니배치크기는 30, 초기 학습률은 0.0001, 학습 횟수(epoch)는 20을 사용하였다. 검증 데이터에 대해 학습된 ResNet-101의 특징벡터를 구하고 그 특징벡터를 2차원까지 t-SNE 방법으로 차원축소 후 가시화하였다. 그림 5.19과 그림 5.20은 냉장고에서 꺼내기의 t-SNE 가시화와 청소기 사용하기의 t-SNE 가시화를 각각 보여준다. 성인은 보라색, 노인은 청록색으로 특징점들을 출력하였고 성인과 노인간의 군집이 있는 것을 볼 수 있다.

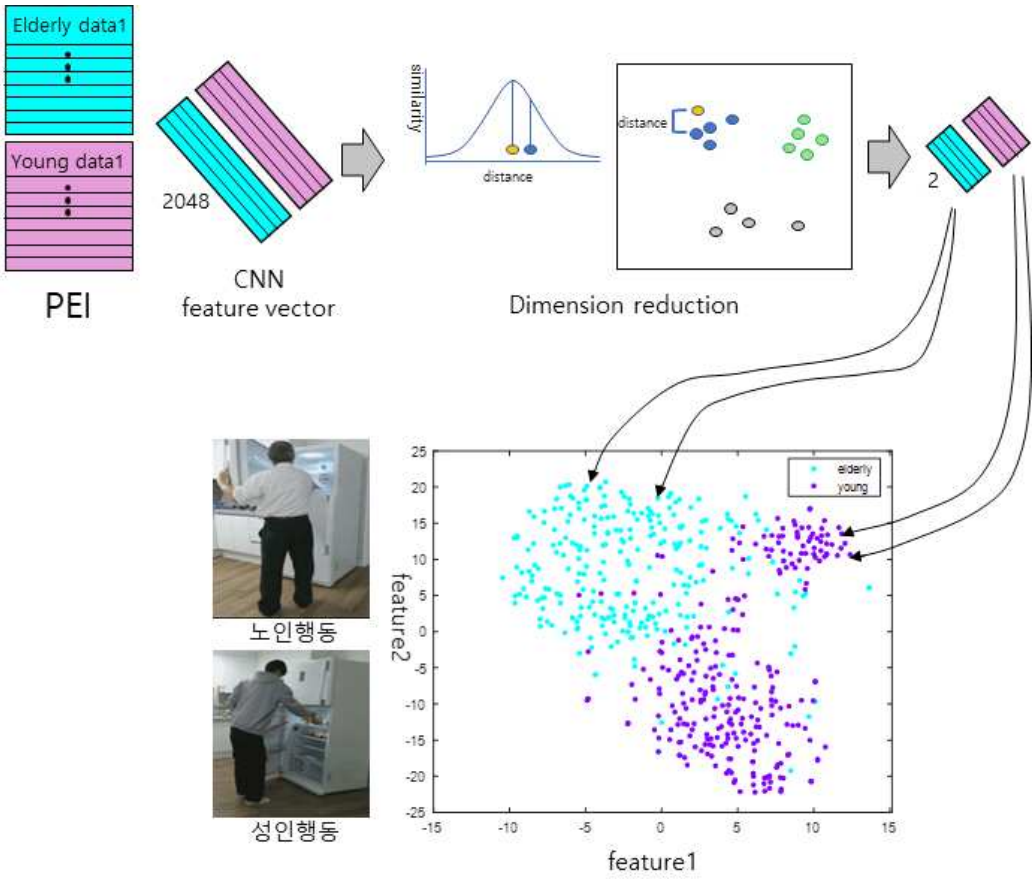


그림 5.19 냉장고에서 꺼내기의 t-SNE 가시화

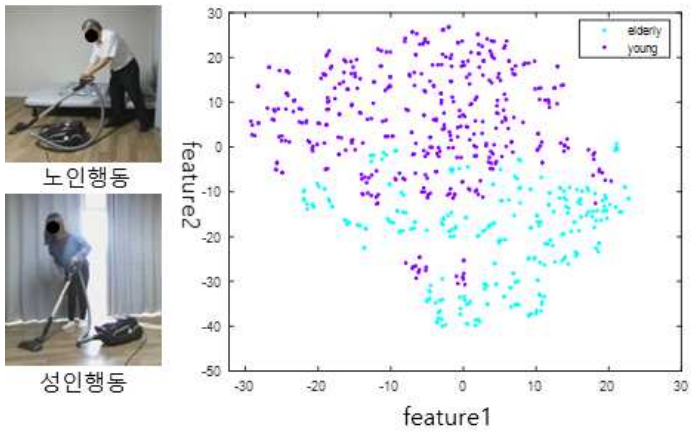


그림 5.20 청소기 사용하기의 t-SNE 가시화

행동특성을 분석하기 위해 설명 가능한 인공지능 기법을 이용할 수 있다. 단일 행동에 대한 스켈레톤 시퀀스를 PEI-T1으로 변환하고 노인과 성인을 분류하도록 ResNet-101을 학습시킨다. 그 후 ResNet-101이 입력의 어디를 보고 출력을 낸 것인지 확인하기 위해 Grad-CAM으로 히트맵을 구한다. 히트맵은 따뜻한 색(붉은색)과 차가운 색(푸른색)으로 표시되며 따뜻한 색이 분류 결과에 주로 기여되고 있음을 의미한다. PEI-T1을 입력으로 하였기 때문에 데이터의 분석이 어려우므로 PEI-T1을 다시 스켈레톤으로 복구하여 히트맵을 표시한 히트궤적을 구한다. 노인과 성인을 분류하는 모델에 대한 히트궤적이므로 노인에 대한 히트궤적은 노인의 특성을 보일 때 성인에 대한 히트궤적은 성인의 특성을 보일 때 단편적으로 히트된다.

그림 5.21은 노인이 포크로 음식을 집어먹는 행동에 대한 히트궤적을 보여준다. 상단의 히트궤적과 하단의 히트궤적이 오른손(초록색)부분에 한하여 비슷한 결과를 보여주지만 왼손(파란색)은 하단의 히트궤적에서만 나타나고 있다. 상단의 히트궤적에서 왼손이 무릎과 떨어져 검출된 것에 반해 하단의 왼손은 무릎과 맞닿아 검출되었고 그 차이로 하단에서만 히트되었기 때문에 일반적 노인은 오른손으로 포크를 사용하면서 왼손을 무릎에 가만히 올려놓은 경향이 있는 것으로 볼 수 있다. 반면 무릎으로부터 허공에 손을 정체해두는 경향은 없는 것으로 볼 수 있다.

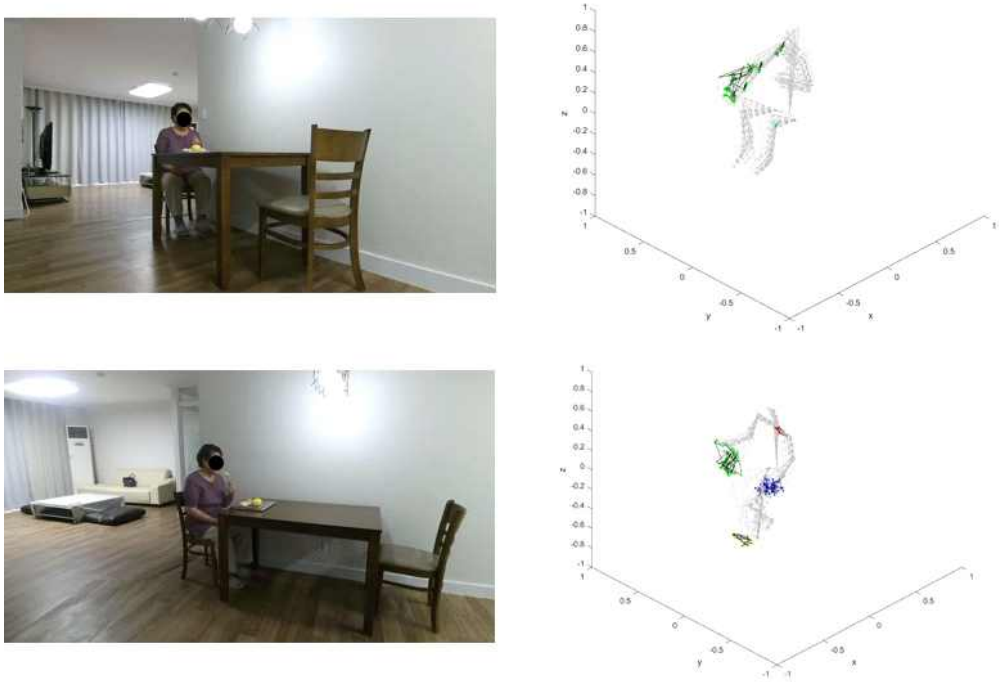


그림 5.21 비슷한 행동에서의 히트궤적 비교

그림 5.22은 노인이 수저로 음식을 떠먹는 행동을 보여준다. 히트궤적에서 왼손이 오른손을 따라가면서 히트된 것을 보아 오른손은 수저를 사용하고 왼손은 음식물을 흘리지 않기 위해 손으로 받치는 경향이 있는 것으로 볼 수 있다.

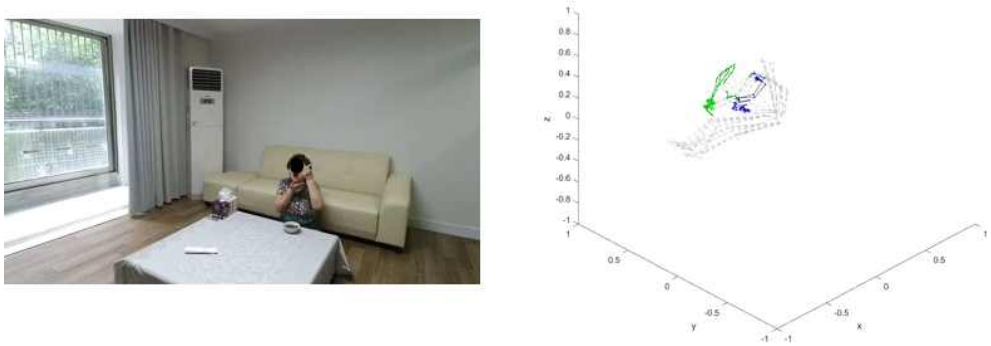


그림 5.22 RGB비디오와 히트궤적 비교(사례1)

그림 5.23은 노인이 포크로 과일을 집어먹는 행동을 보여준다. 히트궤적에서 왼손, 오른손, 목(빨간색)이 히트된 것을 확인 할 수 있다. 왼손은 포크를 사용하면서 오른손은 접시를 잡고 음식이 입 근처에 왔을 때 머리를 앞으로 살짝 내밀며 먹는 경향이 있는 것으로 볼 수 있다.

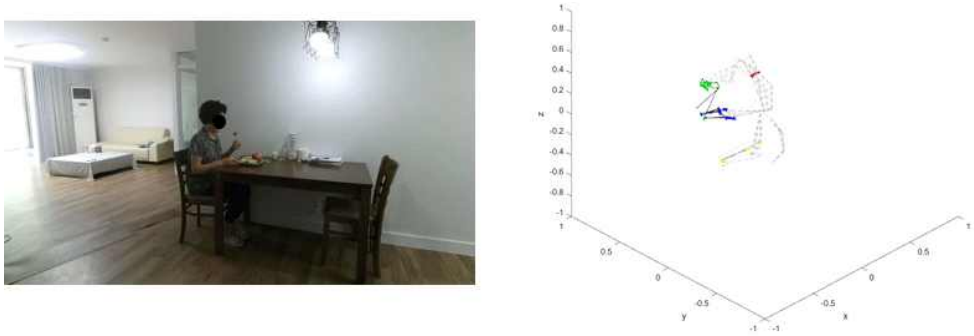


그림 5.23 RGB와 히트궤적 비교(사례2)

그림 5.24은 음식 먹는 행동에 대해 노인과 성인을 비교한 것을 보여준다. 왼쪽 열은 노인이고 오른쪽 열은 성인이다. 노인은 양손과 목에서 주로 히트된 것으로부터 음식을 먹을 때 양손과 목을 주로 활용하는 경향을 볼 수 있지만, 성인은 손, 목과 더불어 팔꿈치, 어깨도 히트되어 상반신을 노인보다 더 활용하며 움직이는 경향이 있는 것으로 볼 수 있다.

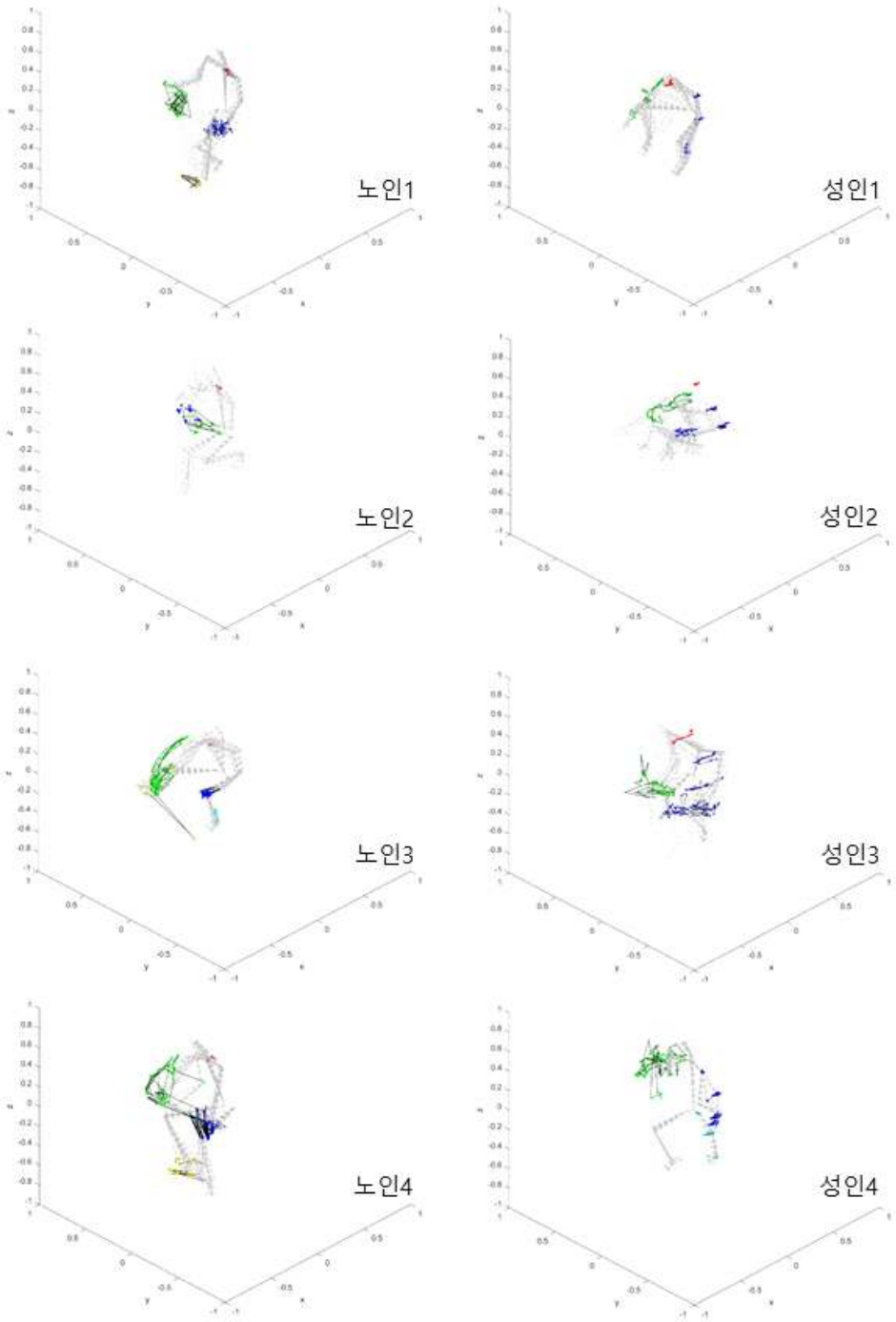


그림 5.24 음식먹는 행동의 노인과 성인 비교

6장 결론

본 논문은 앙상블 및 관심영역기반 심층신경망을 이용한 행동인식과 행동특성 분석을 수행하였다. 비디오기반 행동인식은 대상 사람이 어떤 행동을 하고 있는지 디지털 데이터 처리를 통해 자동으로 알아내는 기술로서 비디오기반 자동 범죄감시, 자동 스포츠 비디오 분석, 실버로봇의 상황인지 등에 응용할 수 있다. 특히 사회의 고령화에 따른 노인 돌봄 문제를 해결하기 위해 실버로봇의 필요성이 증가하면서 그에 핵심 기술로서 행동인식 연구도 중요성이 커지고 있다. 행동인식 데이터는 주로 이미지와 스켈레톤으로 구성되어 특성이 서로 다른 데이터의 분석을 결합하면 더 좋은 인식 성능을 기대할 수 있다. 또한 행동인식의 이미지 데이터는 공간 정보뿐만 아니라 시퀀스로 구성되어 시간정보도 가지고 있다. 그래서 공간정보 또는 시간정보에 각각 최적의 구조로 분석하고 결합함으로서 더 좋은 성능을 기대할 수 있다. 행동인식에서 중요한 정보는 행동을 수행하는 사람 그 자체로 주변 잡음을 제거하고 사람에 관심영역을 두어 신경망을 학습시키면 전체 영역을 학습할 때보다 행동자체에 집중하여 특징분석이 가능하다. 또한 사람은 동물과 다르게 행동을 수행하는데 도구를 사용하므로 손의 물체(hand-object)에 관심영역을 두어 신경망을 학습시키면 도구정보에 집중하여 특징분석이 가능하다. 이러한 관심영역들에 대해 집중하여 학습된 모델들의 정보를 결합하면 더 좋은 성능을 기대할 수 있다. 나이에 따라 신체적 조건이 바뀌기 때문에 행위자의 연령에 따라 데이터의 특성에 차이를 보인다. 이러한 차이에서의 행동특성을 분석하기 위해 설명가능한 인공지능 기법을 이용할 수 있다. 실험을 위해 사용한 데이터셋은 ETRI-Activity3D로 50명의 노인과 50명의 성인의 일상적인 55개 행동에 대한 컬러 이미지, 스켈레톤, 뎁스이미지들을 포함하고 있다. 실험결과로서 제안한 모델인 RGB-S기반 3-스트림 앙상블모델과 관심영역(ROI)기반 앙상블모델이 다른 행동인식 방법들보다 최소 2.6%에서 최대 20.97% 더 성능이 개선되었다. 또한 설명 가능한 인공지능 기법을 통해 스켈레톤 정보로부터 히트궤적을 구하고 RGB비디오와 비교분석을 수행하였다. 향후 노인의 행동 특성을 인지과학적으로 더 명확히 나타낼 수 있는 방법을 연구하고 그 방법을 통해 분석된 행동특성을 딥러닝기반 행동인식 모델의 성능향상에 적용시킬 수 있는 방법을 연구할 계획이다.

참고문헌

[1] 정환수, “한국사회 노인문제에 관한 철학적 고찰”, 새한철학회 논문집, 제 71권, 제 1호, pp. 335-354, 2013.

[2] 김남숙, 사상, “한일중의 고령시대 노인부양에 대한 효문화적 분석”, 한국일본근대학회, 제 66권, 제 0호, pp. 173-193, 2019.

[3] 김재경, “현대사회에 노인문제와 노인인권에 관한 고찰”, 사회복지경영학회, 제 1권, 제 1호, pp. 1-18, 2014.

[4] 한경혜, “농촌노인의 사적 부조”, 한국지역사회생활과학회 학술대회 자료집, pp. 28-35, 1994.

[5] 조명희, “치매노인 부양가족의 부양부담과 개선방안”, 한국노인복지학회, 제 9권, 제 1호, pp. 33-65, 2000.

[6] 이경자, “치매노인의 간호 문제와 돌보는 가족원의 부담감에 관한 연구”, 한국노년학회, 제 15권, 제 2호, pp. 30-51, 1995.

[7] Anastasia K. Ostrowski, Daniella DiPaola, Erin Partridge, Hae Won Park, and Cynthia Breazeal, “Older adults living with social robots”, IEEE Robotics & Automation Magazine, pp. 59-70, 2019.

[8] S. H. Hosseini, K. M. Hoher, “Personal care robots for older adults: an overview”, Asian Social Science, Vol. 13, No. 1, pp. 11-19, 2017.

[9] Joost Broekens, Marcel Heerink, Henk Rosendal, “Assistive social robots in delderly care: a review”, Spring, Vol. 8, No. 2, pp. 94-103, 2009.

[10] Nico Sun, Erfu Yang, Jonathan Corney, Yi Chen, Zeli Ma, “A review of high-level robot functionality for elderly care”, 24th International Conference on Automation and Computing, United Kingdom, 2018.

[11] 김미경, 차의영, “스켈레톤 벡터 정보와 RNN 학습을 이용한 행동인식 알고리즘”, 방송공학회논문지, 제 23권, 제 5호, pp. 598-605, 2018.

[12] 장재영, 홍성문, 손다미, 유호진, 안형우, “지능형 행동인식 기술을 이

용한 실시간 동영상 감시 시스템 개발”, 한국인터넷방송통신학회 논문지, 제 19권, 제 2호, pp. 161-168, 2019.

[13] 정제순, “노화에 따른 골밀도와 신체 수행력의 변화”, 한국체육학회지, Vol. 47, No. 1, pp. 489-499, 2008.

[14] S. Khosla, “Pathogenesis of age-related bone loss in humans”, The Journals of Gerontology, Vol. 68, No. 10, pp. 1226-1235, 2012.

[15] 나보람, 오봉석, “노화와 하체 근력”, 한국노년학연구회, 제 29권, 제 1호, 2020년.

[16] 김재민, 성정섭, 서은선, 고은경, 이석주, 유근창, “노인성 변화에 따른 안구와 해부생리학적 고찰”, 한국안광학회, 제 9권, 제 1호, pp. 135-143, 2004년.

[17] S. Rosemann, C. M. Thiel, “The effect of age-related hearing loss and listening effort on resting state connectivity”, Scientific Reports, Vol. 9, No. 2337, 2019.

[18] R. Peters, “Aging and the brain”, Postgraduate Medical Journal, Vol. 82, No. 964, pp. 84-88, 2006.

[19] “Global medical knowledge”, https://www.msdmanuals.com/

[20] S.Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn", IEEE Conf. Computer Vision and Pattern Recognition, 2018.

[21] H. Wang, and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection", IEEE Trans. Image Processing, Vol. 27, No. 9, 2018, pp. 4382-4394.

[22] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks", IEEE Int. Conf. Multimedia &Expo Workshops, 2017.

[23] S. Yan, X. Yuanjun and L. Dahua, "Spatial temporal graph convolutional networks for skeleton-based action recognition", AAAI conf. Artificial Intelligence, 2018.

[24] Y. H. Wen, L. Gao, H.Fu, F. L. Zhang, and S. Xia, "Graph CNNs with

motif and variable temporal block for skeleton-based action recognition", AAAI Conf. Artificial Intelligence, 2019.

[25] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition", IEEE Signal Processing Letters, Vol. 25, No. 7, 2018, pp. 1044-1048.

[26] C. Xie, C. Li, B. Zhang, C. Chen, J. Han, C. Zou, and J. Liu "Memory attention networks for skeleton-based action recognition", Int. Joint Conf. Artificial Intelligence, 2018.

[27] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation", Int. Joint Conf. Artificial Intelligence, 2018.

[28] A. Kamel, B. Sheng, P. Yang, P. Li, "Deep convolutional neural networks for human action recognition using depth maps and postures", IEEE Transaction on Systems, Man, and Cybernetics: Systems, Vol. 49, No. 9, pp. 1806-1819, 2019.

[29] W. Du, Y. Wang, Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in video", IEEE Transactions on Image Processing, Vol. 27, No. 3, pp. 1347-1360, 2018.

[30] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, "Spatio-temporal attention-based lstm networks for 3d action recognition and detection", IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3459-3471, 2018.

[31] Z. Yang, Y. Li, J. Yang, J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 29, No. 8, pp. 2405-2415, 2019.

[32] H. Sang, Z. Zhao, D. He, "Two-level attention model based video action recognition network", IEEE Access, Vol. 7, pp. 118388-118401, 2019.

[33] J. Yu, H. Gao, W. Yang, W. Chin, N. Kubota, Z. Ju, "A discriminative deep model with feature fusion and temporal attention for human action recognition", IEEE Access, Vol. 8, pp. 43243-43255, 2020.

[34] D. J. Moore, I. A. Essa, M. H. Hayes, "Exploiting human actions and object context for recognition tasks", Proceedings of the 7th IEEE International Conference on Computer Vision ", Greece, 1999.

[35] M. Saitou, A. Kojima, T. Kitahashi, K. Fukunaga, "A dynamic recognition of human actions and related objects", Proceedings of the First International Conference on Innovative Computing, Information and Control, China, 2006.

[36] Y. Gu, W. Sheng, Y. Ou, M. Liu, S. Zhang, "Human action recognition with contextual constraints using a rgb-d sensor", Proceeding of the IEEE International Conference on Robotics and Biomimetics, Chian, 2013.

[37] A. Rosenfeld, S. Ullman, "Hand-object interaction and precise localization in transitive action recognition", 13th Conference on Computer and Robot Vision, Canada, 2016.

[38] A. M. D. Boissiere, R. Noumeir, "Infrared and 3d skeleton feature fusion for rgb-d action recognition", IEEE Access, Vol. 8, pp. 168297-168308, 2020.

[39] G. Liu, J. Qian, F. Wen, X. Zhu, R. Ying, P. Liu, "Action recognition based on 3d skeleton and rgb frame fusion", IEEE/RSJ International Conference on Intelligent Robots and Systems ", pp. 258-264, China, 2019.

[40] Wanqing Li, Zhengyou Zhang Zicheng Liu, "Action recognition based on a bag of 3D points", CVPR Workshops, 2010.

[41] Jaeyong Sung, Colin Ponce, Bart Selman, Ashutosh Saxena, "Human activity detection from RGBD images", In AAAI workshop on Pattern, Activity and Intent Recognition (PAIR), 2011.

[42] B. Ni, G. Wang, and P. Moulin, Rgbd-hudaact: A color-depth video

database for human daily activity recognition“, ICCV Workshops, 2011.

[43] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras”, CVPR, 2012.

[44] Z. Cheng. L. Qin, Y. Ye, Q. Huang, and Q. Tian, “Human daily action analysis with multi-view and color-depth data”, ECCV Workshops, 2012.

[45] H. S. Koppula, R. Gupta, and A. Saxena, “Learning human activities and object aoffrdances from rgb-d videos”, IJRR, 2013.

[46] O. Oreifej, Z. Liu, “HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences”, IEEE Conference on Computer Vision and Pattern Recognition, USA, 2013.

[47] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, “Modeling 4d human-object interactions for event and object recognition, ICCV, 2013.

[48] G. Yu, Z. Liu, and J. Yuan, “Discriminative orderlet mining for real-time recognition of human-object interaction”, ACCV, 2014.

[49] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, “Cross-view action modeling, learning, and recognition”, CVPR, 2014.

[50] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian, “HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition”, ECCV, 2014.

[51] K. Wang, X. Wang. L. Lin, M. Wang, and W. Zuo, “3d human activity recognition with reconfigurable convolutional neural networks”, ACM MM, 2014.

[52] C. Chen, R. Jafari, and N. Kehtarnavaz, “Utd-mhad: A multi-modal dataset for human action recogntion utilizing a depth camera and a wearable inertial sensor”, ICIP, 2015.

[53] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, “Histogram of oriented principal components for cross-view action recognition, TPAMI, 2016.

[54] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, Gang Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis”, Computer

Science, arXiv:1604.02808, 2016

[55] 고병철, “비디오기반 행동인식 연구 동향”, 전자공학회지, 제 44권, 제 8호, pp. 16-22, 2017.

[56] 김무섭, 정치윤, 손종무, 임지연, 정승은, 정현태, 신형철, “스마트폰 기반 행동인식 기술 동향”, 한국전자통신연구원, 제 33권, 제 3호, pp. 89-99, 2018.

[57] Andrej Karpathy, George Toderici, Sanketh Shetty, “Large-scale video classification with convolutional neural networks”, IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725-1732, Ohio USA, 2014.

[58] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici, “Beyond short snippets: deep networks for video classification”, Computer Science, arXiv:1503.08909, 2015.

[59] Karen Simonyan, Andrew Zisserman, “Tow-stream convolutional networks for action recognition in videos”, Computer Science, arXiv:1406.2199, 2014.

[60] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri, “Learning spatiotemporal features with 3D convolution networks, ICCV, pp. 4489-4497, 2015.

[61] Limin Wang, Yu Qiao, Xiaoou Tang, “Action recognition with trajectory-pooled deep-convolutional descriptors”, CVPR, pp. 4305-4314, 2015.

[62] Xiaodong Yang, Pavlo Molchanov, Jan Kautz, “Multilayer and multimodel fusion of deep neural networks for video classification, Proceedings of the 24th ACM international conference on Multimedia, pp. 978-987, 2016.

[63] Shikhar Sharma, Ryan Kiros, Ruslan Salakhutdinov, “Action recognition using visual attention”, ICLR, pp. 1-11, 2016.

[64] Youhui Tian, “Artificial intelligence image recognition method based on convolutional neural network algorithm”, IEEE Access, Vol.

8, pp. 125731-125744, 2020.

[65] Lichao Mou, Pedram Ghamisi, Xiao Xiang Zhou, “Deep recurrent neural networks for hyperspectral image classification”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 55, No. 7, pp. 3639-3655, 2017.

[66] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, Maosong Sun, “Graph neural networks: a review of methods and applications”, Computer Science, arXiv:1812.08434, 2018.

[67] Fabian Horst, Sebastian Lapuschkin, Wojciech Samek, Klaus-Robert Muller, Wolfgang I. Schollhorn, “Explaining the unique nature of individual gait patterns with deep learning”, Scientific Report, Vol. 9, pp. 2391, 2019.

[68] Nanna Notthoff, Peter Reisch, Denis Gerstorf, “Individual characteristics and physical activity in older adults: a systematic review”, Gerontology, Vol. 63, No. 5, pp. 443-459, 2017.

[69] Darcy L. Johannsen, James P. DeLany, Madlyn I. Frisard, Michael A. Welsch, Christina K. Rowley, Xiaobing Fang, S. Michal Jazwinski, Eric Ravussin, “Physical activity in aging: comparison among young, aged, and nonagenarian individuals”, Journal of Applied Physiology, Vol. 105, No. 2, pp. 495-501, 2008.

[70] Ann M. Harris, Lorraine M. Lanningham-Foster, Shelly K. McCrady, and James A. Levine, “Nonexercise movement in elderly compared with young people”, American Journal of Physiology Endocrinology and Metabolism, Vol. 292, No. 4, pp. E1207-E1212, 2007.

[71] Daniel J. Goble, James P. Coxon, Annouchka Van Impe, Jeroen De Vos, Nicole Wenderoth, and Sephan P. Swinnen, “Human Brain Mapping, Vol. 31, No. 8, pp. 1281-1295, 2010.

[72] M. Liu, J. Yuan, “Recognizing human actions as the evolution of pose estimation maps”, IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1159-1168, USA, 2018.

[73] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Communications of the ACM, Vol. 60, No. 6, pp. 84-90, 2017.

[74] I. Sutskever, O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural networks", Computer Science, arXiv:1409.3215, 2014.

[75] K. Hara, H. Karaoka, Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition", Computer Science, arXiv:1708.07632, 2017.

[76] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions", IEEE Conference on Computer Vision and Pattern Recognition, USA, 2015.

[77] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", Computer Science, arXiv:1512.03385, 2015.

[78] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, "Densely connected convolutional networks", Computer Science, arXiv:1608.06993, 2017.

[79] J. Carreira, A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset", Computer Science, arXiv:1705.07750, 2018.

[80] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, "A closer look at spatiotemporal convolutions for action recognition", IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6450-6459, USA, 2018.

[81] Q. Wang, G. Kurillo, F. Ofli, R. Bajcsy, "Evaluation of pose tracking accuracy in the first and second generations of microsoft kinect", International Conference on Healthcare Informatics, pp. 380-389, USA, 2015.

[82] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, Y. A. Sheikh, "OpenPose: realtime multi-person 2d pose estimation using part affinity fields", IEEE Transactions of Pattern Analysis and Machine

Intelligence, 2019.

[83] J. Jang, D. H. Kim, C. Park, M. Jang, J. Lee, J. Kim, “ETRI-activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly”, IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 10990-10997, USA, 2020.

[84] 안재현, XAI 인공지능을 해부하다, 위키북스, 2020.

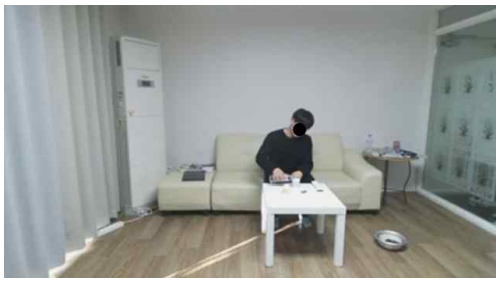
[85] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, “Learning deep features for discriminative localization”, IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921-2929, USA, 2016.

[86] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient based localization”, IEEE Conference on Computer Vision, pp. 618-626, Italy, 2017.

[87] L. V. D. Maaten, G. Hinton, “Visualizing data using t-sne”, Journal of Machine Learning Research, Vol. 9, pp. 2579-2605, 2008.

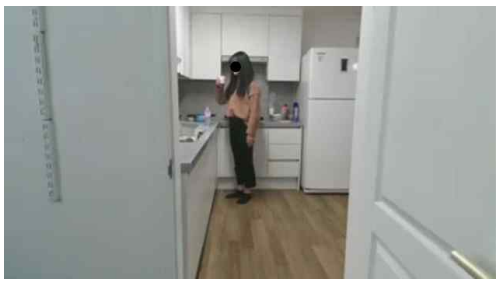
부록

(a) 행동01



(b) 행동02

(c) 행동03



(d) 행동04

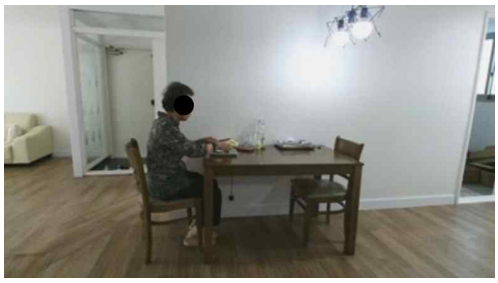
(e) 행동05

(f) 행동06

(g) 행동07

(h) 행동08

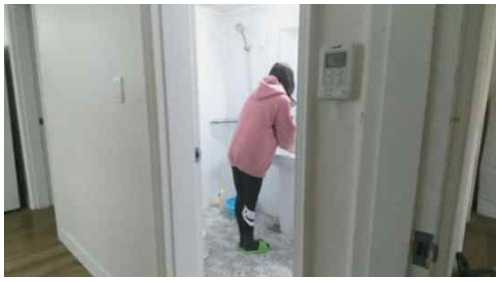
부록 1 ETRI-Activity3D 행동01~08



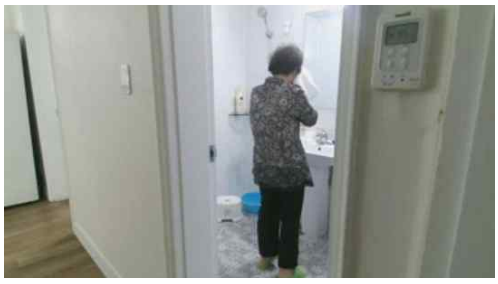
(a) 행동09

(b) 행동10

(c) 행동11



(d) 행동12



(e) 행동13

(f) 행동14

(g) 행동15

(h) 행동16

부록 2 ETRI-Activity3D 행동09~16

(a) 행동17

(b) 행동18

(c) 행동19

(d) 행동20

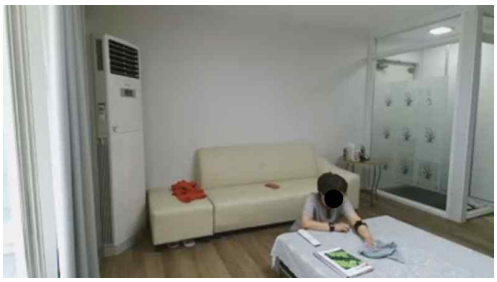
(e) 행동21

(f) 행동22

(g) 행동23

(h) 행동24

부록 3 ETRI-Activity3D 행동17~24



(a) 행동25

(b) 행동26

(c) 행동27

(d) 행동28

(e) 행동29

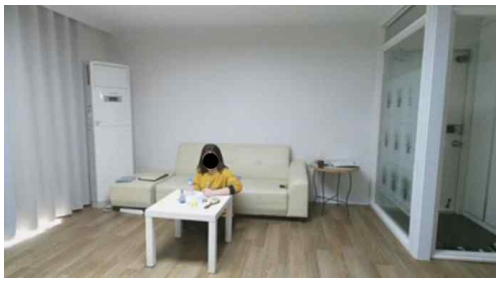
(f) 행동30

(g) 행동31

(h) 행동32

부록 4 ETRI-Activity3D 행동25~32

(a) 행동33



(b) 행동34

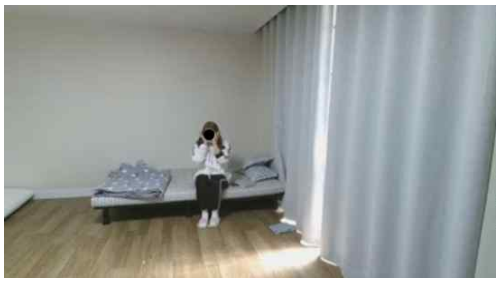
(c) 행동35

(d) 행동35

(e) 행동37

(f) 행동38

(g) 행동39



(h) 행동40

부록 5 ETRI-Activity3D 행동33~40

(a) 행동41

(b) 행동42

(c) 행동43

(d) 행동44

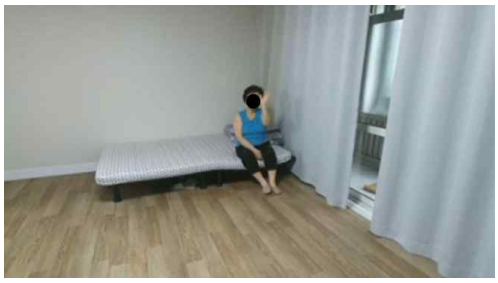
(e) 행동45

(f) 행동46

(g) 행동47

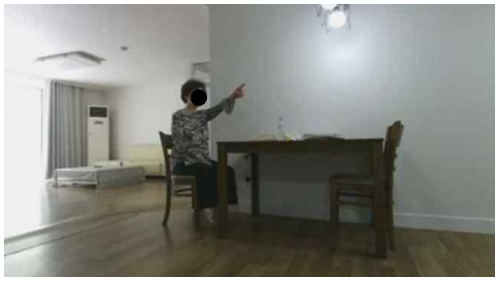
(h) 행동48

부록 6 ETRI-Activity3D 행동41~48



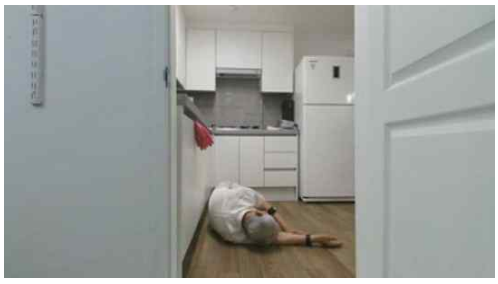
(a) 행동49

(b) 행동50



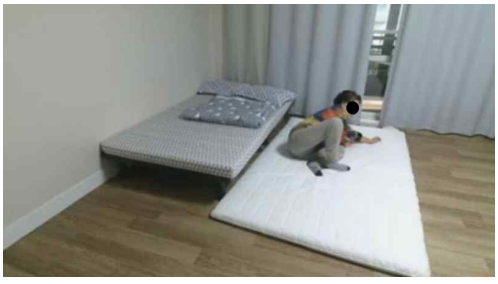
(c) 행동51

(d) 행동52



(e) 행동53

(f) 행동54



(g) 행동55

부록 7 ETRI-Activity3D 행동49~55