JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 12, pp. 33-40, December 2023
https://doi.org/10.9708/jksci.2023.28.12.033

A Study on Non-Contact Care Robot System through Deep Learning

Hyun-Sik Ham*, Sae Jun Ko*

*Researcher, ZIOVISION, Chuncheon, Korea
*Researcher, ZIOVISION, Chuncheon, Korea

[Abstract]

As South Korea enters the realm of an super-aging society, the demand for elderly welfare services has been steadily rising. However, the current shortage of welfare personnel has emerged as a social issue. To address this challenge, there is active research underway on elderly care robots designed to mitigate the social isolation of the elderly and provide emergency contact capabilities in critical situations. Nonetheless, these functionalities require direct user contact, which represents a limitation of conventional elderly care robots. In this paper, we propose a solution to overcome these challenges by introducing a care robot system capable of interacting with users without the need for direct physical contact. This system leverages commercialized elderly care robots and cameras. We have equipped the care robot with an edge device that incorporates facial expression recognition and action recognition models. The models were trained and validated using public available data. Experimental results demonstrate high accuracy rates, with facial expression recognition achieving 96.5% accuracy and action recognition reaching 90.9%. Furthermore, the inference times for these processes are 50ms and 350ms, respectively. These findings affirm that our proposed system offers efficient and accurate facial and action recognition, enabling seamless interaction even in non-contact situations.

▸ **Key words**: Action Recognition, Care Robot, Deep Learning, Edge Device, Face Detection, Facial Expression Recognition

[요 약]

한국이 초고령사회로 진입하면서 노인 복지에 대한 필요성이 증가하고 있으나 현재 복지 인력 부족이 사회문제로 대두되고 있다. 이에 대한 해결책으로 노인의 사회적 고립감 완화와 위급 상황 시 비상 연락 등의 기능을 하는 노인 돌봄 로봇이 활발히 연구되고 있다. 하지만 이러한 기능들은 사용자의 접촉이 있어야만 작동하여 기존 노인 돌봄 로봇의 한계점으로 자리 잡고 있다. 본 논문에서는 기존의 문제를 해결하기 위해 상용화된 노인 돌봄 로봇과 카메라를 통해 직접적인 접촉 없이도 사용자와 상호작용할 수 있는 돌봄 로봇 시스템을 제안한다. 돌봄 로봇에 연결된 엣지 디바이스에 표정 인식 모델과 행동 인식 모델을 탑재하였고, 공공데이터를 통해 모델의 학습 및 성능검증을 진행했다. 실험 결과를 통해 표정 인식과 행동 인식의 성능이 각각 정확도 96.5%, 90.9%인 것을 확인할 수 있으며, 수행 시간의 경우에는 각각 50ms, 350ms인 것을 확인할 수 있다. 해당 결과는 제안한 시스템의 표정 및 행동 인식 정확도가 높고 추론 시간이 효율적임을 확인하며, 이는 비접촉 상황에서도 원활한 상호작용을 가능하게 한다.

▸ **주제어**: 돌봄 로봇, 딥러닝, 얼굴 검출, 엣지 디바이스, 표정 인식, 행동 인식

• First Author: Hyun-Sik Ham, Corresponding Author: Hyun-Sik Ham
*Hyun-Sik Ham (ham3868@ziovision.co.kr), ZIOVISION
*Sae Jun Ko (a2009mys@ziovision.co.kr), ZIOVISION
• Received: 2023. 11. 28, Revised: 2023. 12. 14, Accepted: 2023. 12. 19.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

현재 대한민국은 인구 통계 집계를 시작한 뒤 최초로 사망자 수가 출생자 수를 넘어서 인구가 감소하고 있다. 통계청은 장래인구추계를 통해 그림 1에서 볼 수 있듯이 2020년 이후 10년 동안 생산가능인구가 357만 명 감소하고 고령인구는 490만 명이 증가한다고 예측했다[1]. 또한, 65세 이상 고령인구의 비율은 2025년에 20.6%에서 2050년 40%를 초과할 것으로 전망되고 있다. 저출산으로 인한 생산가능인구 감소와 노인층의 증가로 한국의 초고령사회 진입은 불가피한 현상이 되었다.

초고령사회로 진입함에 따라 사회 복지사에 대한 요구는 날로 증가하고 있다. 사회 복지사의 자택 방문과 다양한 사회 활동으로 거동이 불편한 노인을 돌봐주거나 노인의 사회적 고립감을 해소해줄 수 있어 초고령사회에서 사회 복지사는 각광받고 있는 직업이 되었다. 하지만 늘어가는 노인 인구를 담당할 수 있는 인력은 한정되어 있어 제한된 수의 노인들에게만 복지서비스를 제공할 수밖에 없다. 이는 복지의 질이 하락하고 접근성이 떨어져 복지 사각지대를 형성해 결국 노인의 가족이나 사회의 부담이 증가하는 사회문제가 되고 있다. 앞선 문제를 해결하고 노인 복지를 향상하기 위해 지원금과 복지시설의 확충 및 인력의 증원과 같은 해결책이 있으나 이는 사회적으로 막대한 비용이 발생한다는 문제가 있다. 최근에는 인력 부족 문제를 보조할 하나의 수단으로 돌봄 로봇에 관한 연구가 진행되고 있다. 돌봄 로봇은 의료 및 복지 분야에서 거동이 불편한 환자를 위해 이동, 식사 그리고 배변 등을 보조하거나 사회 복지사를 도와 노인에게 외로움 같은 사회적 고립감 해소 등의 역할을 한다.

노인 돌봄서비스를 제공하는 돌봄 로봇은 주로 음성 및 센서 기반으로 동작하고 있어 사용자와의 접촉 및 대화와 같은 직접적인 상호작용이 필수이다. 이러한 돌봄 로봇은 사용자가 직접 만지고 있지 않을 때 어떠한 대처도 불가능하다는 단점이 존재한다. 특히, 갑자기 쓰러지는 것과 같은 모든 비접촉 위급 상황에 대처하지 못한다. 비접촉 문제를 해결하기 위해서는 센서로 획득할 수 없는 시각 정보를 얻을 수 있는 영상 장치가 필요하다. 영상 장치를 통해 사용자의 표정에서 감정을 인식하고, 무슨 행동을 취하는지를 인지할 수 있다면 사용자의 요구를 파악할 수 있어 사용자의 만족도가 높아질 수 있으며 기존의 문제점인 비접촉 위급 상황에 대처할 수 있다. 하지만 현재 한국에서 상용화된 노인 돌봄 로봇 중에서는 영상 기능을 제공하는 제품은 존재하지 않는다.

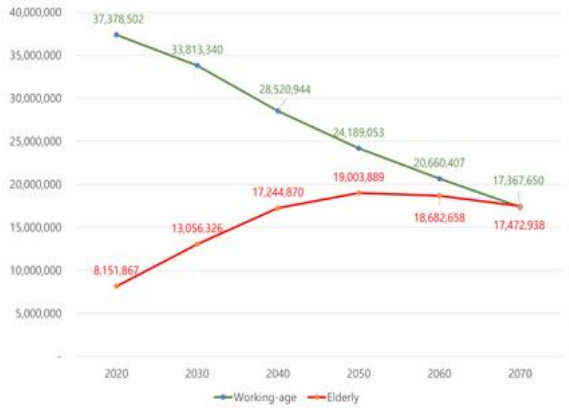

Fig. 1. Trends in Korea's Working-Age and Elderly Population

본 논문에서는 기존 돌봄 로봇에 카메라를 설치해 표정과 행동으로 사용자의 감정을 인식하는 시스템을 제안한다. 일반 PC에서 표정을 통해 화남, 슬픔 등의 심리 상태를 파악하는 표정 인식(Facial Expression Recognition, FER) 모델과 약 먹기와 TV 시청 같은 생활 습관 및 쓰러짐을 감지하는 행동 인식(Action Recognition) 모델을 학습해 성능을 평가했으며, 모델들을 돌봄 로봇으로 이식하여 추론 속도를 측정하여 제안 시스템의 적합성을 검증하였다.

II. Related works

1. Care Robot

전 세계적으로 고령화 사회 문제가 대두되면서 돌봄 로봇에 관한 연구가 폭넓게 진행되고 있다. 특히 최근 딥러닝의 발전과 모바일 기기의 성능이 높아짐에 따라 다양한 딥러닝 모델이 빠르게 연구되고 있으며, 다양한 기능을 탑재한 로봇들이 출시되고 있다.

기관에서 사용되는 대표적인 돌봄 로봇으로는 로봇팔 디자인의 보조 돌봄 로봇이 있다[2]. Lio라고 불리는 로봇의 크기는 팔 관절을 모두 폈을 때 약 1.6m이고 로봇팔 아래로 바퀴가 달린 플랫폼이 있어 이동할 수 있는 로봇이다. 4개의 다른 기기를 사용해 수집한 정보를 처리하며 서로 연결된 이더넷으로 정보를 주고받는다. 자율 주행 기능을 탑재하였으며 주행 중 충돌을 피하고자 어안렌즈 카메라와 LiDAR 그리고 깊이 카메라를 설치했다. AI 알고리즘의 경우에는 물체를 탐색하고 인지하기 위한 객체 탐지 알고리즘과 사람 인식을 위한 얼굴 탐지 등을 사용하며, 이러한 하드웨어와 AI 알고리즘을 통해 Lio는 거동이 불편해

휠체어를 탄 환자 대상으로 바닥이나 선반 위의 물체를 대신 집어주는 역할을 한다.

가정에서 사용할 수 있도록 연구한 사례로 DORI라는 돌봄 로봇이 있다. 곰 인형 모습의 돌봄 로봇인 DORI는 노인 대상 서비스를 목표로 연구가 진행됐다[3]. 곰 인형이 움직일 수 있는 플랫폼 위에 놓여있고 LiDAR, 카메라, 터치 센서 등이 설치되어 각종 정보를 수집한다. 대표적인 기능으로 카메라와 자세 추정 알고리즘인 BlazePose를 이용해 사용자의 일어섬, 앉음, 쓰러짐을 판단한다[4]. 또한, 사용자를 따라다니며 정보를 수집하는 관찰 모드를 음성으로 켜고 끌 수 있게 했다.

Lio는 의료 기관 내에서 거동이 불편한 환자와 바쁜 의료 인력을 보조해줄 수 있는 돌봄 로봇이지만 상대적으로 큰 크기로 가정에서 쓸 수 있는 노인 돌봄 로봇에 적합하지 않다. 반면에 DORI는 가정에서 쓸 수 있는 크기의 로봇이며 외관도 친근함을 느낄 수 있는 곰 인형 형태이지만 아직 연구 단계라 상용화되지 않았다. 본 논문에서는 이미 상용화된 돌봄 로봇에 영상 모듈을 추가하는 방식으로 위의 문제를 해결하고, 사용자의 감정 및 행동을 예측 모델을 탑재한 돌봄 로봇을 만들기 위한 시스템을 제안한다.

2. Facial Expression Recognition

표정 인식은 검출한 대상의 얼굴을 통해 감정을 분류하는 기술을 의미한다. 표정 인식의 대표적인 데이터셋으로 FER-2013 데이터셋이 있다. FER-2013은 화남, 공포, 행복, 슬픔 등을 포함한 7개의 감정 클래스가 있으며 총 35,888장의 데이터로 구성되어 있다. FER-2013 데이터셋을 이용한 표정 인식 연구 중에 VGG를 사용하여 학습시킨 연구가 있다[5][6]. 분류 네트워크인 VGG을 통해 학습 데이터셋을 학습하고 검증용 데이터셋과 합쳐 작은 학습률(Learning rate)로 다시 학습시켰다. 그 결과 기존의 FER-2013 학습 모델보다 좋은 성능인 정확도 73.28%를 달성하였다.

다른 연구로는 경량화된 모델 중 하나인 MobileNetV1으로 표정 인식을 시도한 방법도 있다[7]. 해당 방법에서는 표정 인식에 필수적인 얼굴 표현 특징을 잘 추출하기 위해 가벼운 어텐션(Attention) 모듈을 사용했으며, 다른 클래스의 특징 간 거리는 멀게 하고 같은 클래스의 특징은 가깝게 학습시키기 위해 softmax loss와 center loss를 사용하였다. FERPlus와 RAF-DB(Real-world Affective Faces Database) 데이터셋으로 학습 및 테스트를 진행하였다. FERPlus는 기존 FER2013 데이터셋의 레이블을 10명이 수정한 데이터셋이며 RAF-DB는 40명이 7개의 표정 클래스로 레이블링을 진행한 29,672장으로 구성된 데이터셋이다. 실험 결과 각 데이터셋에 대해 정확도는 각각 88.11%와 84.49%로 연구에서 제안한 모델이 다른 경량화 모델인 MobileNetV2와 MobileNetV3보다 나은 성능을 보였다.

3. Action Recognition

행동 인식은 주로 연속된 이미지로 구성된 영상을 입력으로 받아 사람의 행동을 분류한다. 이상 행동을 인식하는 사례로 자세 추정(Pose Estimation)과 그래프 컨볼루션 오토인코더(Graph Convolutional Autoencoder)를 통한 이상 행동 탐지 연구가 있다[8]. 자세 추정 알고리즘으로 VGG-19 백본의 사용과 객체 탐지 알고리즘를 사용하지 않은 bottom-up 방식의 알고리즘 채용으로 추정 시간을 줄여 실시간 탐지를 목표로 개발된 OpenPose를 사용하였다[9]. 시간 특징을 학습하기 위해 한 영상의 12개 프레임을 모아 학습했으며, 추정한 스켈레톤으로 행동 특징을 추출한 뒤 디코더로 다시 복원해 추정한 스켈레톤과 복원한 스켈레톤 사이의 복원 오차를 계산해 학습하였다. 오차 임곗값을 설정해 복원 오차가 임곗값 이하면 정상, 이상이면 이상 행동으로 예측하였다. 이상 행동은 쓰러짐, 싸움, 서기, 앉기, 눕기 5가지를 선정해 분류했다. 임곗값을 바꿔가며 성능을 측정해 그린 그래프의 AUC(Area Under the Curve)를 성능 지표로 하여 적층 오토인코더나 LSTM-오토인코더를 활용한 다른 비지도학습 모델과 비교했을 때, 5개 행동 중 4개에서 제안한 방법이 가장 좋은 성능을 달성했다.

깊이 카메라로 촬영한 노인 행동 분류 연구도 있다[10]. 요양센터의 방에 깊이 카메라를 설치해 초당 1프레임으로 깊이 정보를 받아 데이터셋을 구성했다. 사람이 포함되지 않은 30프레임 영상의 평균으로 배경 이미지를 생성하고 나중에 촬영되는 이미지와 배경 간의 차영상을 통해 사람과 휠체어의 위치를 탐지한다. 그리고 깊이 이미지 내에서 침대와 바닥을 관심 영역으로 설정해 사람의 위치에 따라 쓰러짐, 침대에 누움, 휠체어에 앉음 등의 여덟 행동을 분류했다. 분류 성능은 정확도 93.7%로 깊이 카메라로도 행동 분류가 가능함을 보였다.

자세 추정 알고리즘을 행동 인식에 이용하는 것은 성능면에서 개선이 있지만 계산 비용이 많이 필요해 자원이 제한된 엣지 디바이스에서는 매우 느린 추론 속도를 보여 제안하는 돌봄 로봇 시스템에선 적합하지 않다. 또한, 깊이 카메라는 기본적으로 모듈이 커서 기존의 시판된 돌봄 로봇에 부착하기 어렵다는 문제가 있어 본 논문에서는 일반

적인 RGB 카메라만을 이용하여 행동 인식을 진행하였다. 제안한 시스템에서는 모바일 디바이스와 RGB 카메라만을 이용하였고, 속도와 성능 면에서 우수한 비전 트랜스포머(Vision Transformer) 모델을 활용해 행동 인식 모델을 구현하였다[11].

III. Method

1. Facial Expression Recognition

1.1 Dataset of Facial Expression Recognition

본 논문에서 사용한 표정 인식 데이터셋은 한국지능정보사회진흥원에서 주관하는 AI Hub의 한국인 감정인식을 위한 복합 영상 데이터셋을 사용하였다[12]. 국외 공개 데이터셋의 경우 AI Hub 데이터셋 대비 많은 양을 보유하고 있지만, 백인과 흑인의 비율이 동양인에 비해 많은 편으로 한국인 노인을 대상으로 하는 국내 노인 돌봄 로봇에 활용하기엔 적합하지 않다고 판단하여 AI Hub 한국인 데이터셋을 사용하여 학습을 진행하였다.

AI Hub의 한국인 표정 인식 데이터셋은 기쁨, 당황, 분노, 불안, 상처, 슬픔, 중립의 일곱 가지의 감정을 수집했고, 약 48만 건의 데이터로 구성되어 있다. 어노테이션은 사람의 감정과 얼굴 위치에 대한 경계 상자 그리고 배경 정보를 포함하고 있다. 본 논문에는 기본 정서인 기쁨, 분노, 슬픔, 중립 네 가지의 감정만 학습 및 테스트를 진행하였다. 자세한 데이터 구성은 아래 표 1과 같다. 데이터셋의 전처리는 다음과 같이 진행하였다. 먼저 어노테이션에 있는 얼굴 경계 상자를 활용해 원본 이미지 내에서 얼굴 영역을 추출하였다. 그 후 얼굴 영역을 224×224 크기로 조정한 뒤 모델에 입력으로 사용하였다.

Table 1. Configurations of AI Hub Facial Expression Recognition Dataset

Class	Train	Test
Angry	59,624	7,451
Happy	59,937	7,508
Neutral	59,163	7,381
Sad	59,779	7,469
Total	238,503	29,809

1.2 Face Detection Model

원활한 표정 인식을 위해선 이미지 내의 사람 얼굴을 검출해주는 얼굴 검출 모델이 선행되어야 한다. 구글에서 개발한 모바일 기계학습 SDK인 Google ML Kit는 모바일에서 간편하게 텍스트 인식, 객체 감지 등 기계학습의 다양한 기능을 구현할 수 있으며 그중에서 얼굴 인식(Face Detection) API도 제공하고 있다. 얼굴 인식 API는 얼굴 특징 인식 및 위치 파악 기능으로 구성되며, 카메라에서 영상을 받아 실시간 처리가 가능하여 제안 시스템에서는 해당 API를 통해 얼굴 검출을 진행한다.

1.3 Facial Expression Recognition Model

본 논문에서는 모델의 Width, Depth 그리고 입력 해상도를 다양하게 조절하면서 성능을 최적화한 모델인 EfficientNet을 사용해 표정 인식 모델을 구현하였다[13]. 제한된 성능을 고려하여 MobileNet과 모델 파라미터 수가 크게 차이 나지 않으면서 우수한 성능을 보인 B0 모델을 선정했다. 전 세계 사람의 얼굴을 약 삼백만 장 수집하고 약 9,000개 이상의 클래스로 분류한 데이터셋인 VGGFace2을 통해 학습된 EfficientNet-B0를 사전학습 모델로 가져와 미세조정을 진행했다[14][15].

2. Action Recognition

2.1 Dataset of Action Recognition

행동 인식 데이터셋은 한국전자통신연구원(ETRI)에서 공개한 ETRI-Activity3D를 사용하였다[16]. ETRI-Activity3D 데이터셋은 로봇이 사람에게 노인 대상 휴먼케어 서비스를 제공할 수 있도록 고령자 대상으로 제작한 데이터셋이며, 노인 100명을 대상으로 각기 다른 시점의 8개 카메라를 설치하고 음식 섭취, 여가, 가사, 사람과의 상호작용 등의 총 55개 행동을 녹화한 데이터셋이다.

본 연구에서는 데이터셋의 행동 중 노인의 일상생활과 밀접한 약 먹기(Take Medicine), 리모컨 조작(Remote Control), 쓰러짐(Fallen)을 선정하여 실험을 진행하였다. 특히, 약 먹기는 약을 먹는 행동이 영상 내에서 극히 일부 구간이며 나머지는 특정한 행동을 취하지 않기 때문에 모든 구간을 약 먹기라고 레이블링하는 경우 오탐이 증가해 행동 없음(None) 클래스를 추가하여 해당 문제를 해결하였다.

영상에서 배경이 모델 학습에 미치는 영향을 최대한 배제하기 위해 사람의 관심 영역을 잘라내어 전처리를 진행했다. 관심 영역을 추출한 한 영상을 3초 길이로 나눠 클립을 생성했고 각 클립은 이전 클립의 마지막 0.5초와 겹치게 샘플링했다. 약 먹기와 행동 없음의 경우 한 영상에서 분리돼 리모컨 조작과 쓰러짐에 비해 데이터가 부족하여 모델 성능이 편향되는 모습을 보여주었다. 이에 행동 없음과 약 먹기 학습 데이터를 각각 4배, 7배 단순 복사로 증강하여 학습에 사용하여 해결하였다. 단순 복사에 따라

기존의 데이터 수가 행동 없음 581개에서 2,324개로 증대되었고, 약 먹기 293개에서 2,051개로 증가하였다.

Table 2. Configurations of ETRI-Activity3D Dataset

Class	Train	Val	Test
None	2,324	40	52
Take Medicine	2,051	36	44
Remote Control	2,043	270	320
Fallen	1,818	218	212
Total	8,236	564	628

2.2 Action Recognition Model

행동 인식 모델은 비전 트랜스포머를 통해 학습을 진행하였다. 비전 트랜스포머는 트랜스포머를 컴퓨터 비전에서도 쓸 수 있게 변형한 네트워크로, 포지션 임베딩(Position Embedding)과 query, key, value를 통한 self-attention 연산이 주를 이루는 모델이다. 비전 트랜스포머는 이미지를 패치 단위로 나누어 임베딩하고 변환된 임베딩 벡터로 다른 임베딩 벡터와의 어텐션 연산을 통해 데이터셋의 특징을 학습한다.

입력 이미지의 크기인 224×224에서 패치 크기는 32×32로 설정해 [CLS] 토큰까지 총 50개의 임베딩을 사용하였다. 연구에서 사용한 비전 트랜스포머는 CLIP으로 학습된 가중치를 사용했다. CLIP은 대표적인 멀티모달 모델로 자연어와 이미지로 구성된 대량의 데이터셋에서 학습한다[17]. CLIP은 이미지를 처리하는 비전 인코더와 자연어를 처리하는 텍스트 인코더로 구성되어 있어 이미지와 이미지의 해설 텍스트를 한 쌍으로 입력받아 각 인코더를 통해 이미지 임베딩 벡터와 텍스트 임베딩 벡터를 추출한다. 두 임베딩 벡터 간의 유사도를 구하고 학습에 반영해 비슷한 특징을 지닌 이미지와 텍스트 표현은 가깝게, 다른 특징을 지닌 표현들은 멀게 학습하여 풍부한 특징을 추출하는 모델로 학습된다. CLIP으로 사전학습된 비전 트랜스포머를 최대한 활용하기 위해 사전학습 가중치를 고정하여 ETRI 데이터셋에서 좋은 특징을 추출할 수 있게 했다.

행동을 인식하기 위해서는 공간 정보뿐만 아니라 시간 정보를 학습하는 것이 중요하기 때문에 대다수의 행동 인식 모델은 시간적 정보를 보다 잘 추출할 수 있는 방향으로 발전해왔다. 하지만 데이터의 차원이 시간 축으로 하나 더 추가되면서 연산량과 모델의 크기가 증가한다는 단점이 있다. 본 연구에서는 제한된 디바이스 자원를 활용하기 때문에 연산량과 모델의 크기 증가는 치명적인 요소로 작용할 수 있어 기존 방식과는 다른 방법으로 시간 정보를 학습시켰다. 비전 트랜스포머는 기존처럼 프레임마다 특징을 추출하고 여덟 프레임 분량의 특징이 추출되면 여덟 개의 [CLS] 토큰을 연결해서 Multi-Layer Perceptron(MLP)로 특징 간의 시간 정보를 학습하도록 했다. 학습은 특징 추출기의 역할을 하는 비전 트랜스포머를 제외하고 분류기인 MLP만 진행된다.

3. Care Robot System

본 연구는 국내에서 상용화된 노인 돌봄 로봇 중 효돌이를 기반으로 진행하였다[18]. 기존 효돌이는 인형 내의 센서를 이용해 만지는 부위마다 기능을 다르게 하여 사용자와 상호작용하거나 위급 상황 시 알림 등이 구현되어 있다. 하지만 센서 사용만으로 얻을 수 있는 정보가 한정적이며 비접촉 상황일 때 위급 상황에 대처하지 못한다. 비접촉일 때 얻을 수 있는 대표적인 정보로는 영상 정보가 있으며 행동 인식을 이용한 위급 상황 탐지와 표정을 통한 감정 정보를 포착할 수 있다. 이런 비접촉 시스템 구축을 위해 효돌이에 엣지 디바이스인 KHADAS사의 VIM4을 설치하였고 해당 장치에 RGB 카메라를 부착하여 그림 2처럼 행동 인식 및 표정 인식 모델을 구동하였다[19]. Pytorch로 개발된 모델을 엣지 디바이스에 이식시키기 위해서 Tensorflow-Lite로 변환하여 엣지 디바이스에 이식하였다.

IV. Results

표정 인식과 행동 인식 모델에 대한 성능 테스트는 PC를 통해 진행하였으며, 제안 시스템의 추론 시간은 VIM4 디바이스을 사용하여 측정하였다. 표정 인식과 행동 인식은 각 학습 데이터셋의 테스트셋으로 평가를 진행했다. 모델의 정량적 평가는 Recall, Precision 그리고 F1-Score을 통해 진행했으며 수식은 아래와 같다.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{3}$$

모델 평가를 위한 실험 환경은 표 3과 같다.

Table 3. Environments for Training Facial Expression and Action Recognition Models

CPU	Intel i5-12600 3.3GHz 6 cores
GPU	NVIDIA GeForce RTX 3060 (12GB)
RAM	16GB
OS	Microsoft Windows 11

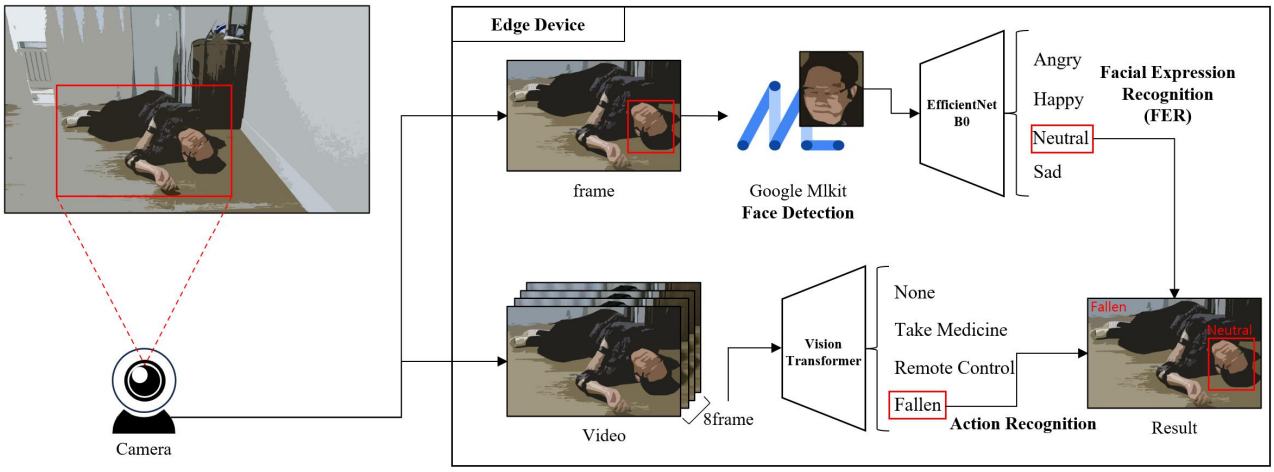

Fig. 2. System of the Proposed Elderly Care Robot

1. Results of Facial Expression Recognition

표정 인식 모델의 학습은 배치 크기(Batch Size) 128로 15에폭(Epoch) 동안 진행되었다. 처음 5에폭은 EfficientNet-B0 모델의 가중치를 고정하고 학습률 1e-3으로 분류기 부분만 학습하였다. 그 후 가중치 고정을 해제하고 학습률을 1e-4로 낮춰 학습을 진행하였다.

표정 인식 결과는 표 4와 같다. 실험 결과는 모든 정량적 평가에서 약 97%의 성능을 보여주고 있다. 특히 기쁨은 다른 클래스와 비교해 입꼬리가 올라가고 눈이 반달 모양이 되는 등의 특징이 도드라지기 때문에 가장 높은 성능을 보여주고 있다. 반대로 슬픔은 눈물을 흘리지 않고 입꼬리를 내리거나 눈썹을 움직여 표현한 방식이 화남 및 무표정과 유사하여 오탐이 발생하였다. 다른 표정들과 비교해 상대적으로 성능이 낮지만 모든 평가에서 약 95% 이상의 성능을 보여줌으로 슬픔도 잘 인식한다는 것을 확인할 수 있다.

Table 4. Results of Facial Expression Recognition Model

Class	Recall	Precision	F1-Score
Angry	0.96	0.96	0.96
Happy	0.98	0.99	0.99
Neutral	0.97	0.96	0.97
Sad	0.95	0.96	0.95
Mean	0.97	0.97	0.97

2. Results of Action Recognition

행동 인식 모델의 배치 크기는 8, 학습률은 1e-3으로 설정하여 100에폭동안 학습했다. 학습 시 검증 데이터셋에서의 정확도가 가장 높은 모델을 성능 평가 및 돌봄 로봇에 사용했다. 표 5는 행동 인식 결과이다. 모델은 ETRI 데이터셋에 대해 약 90.9%의 정확도를 보였다. 쓰러짐은 사람이 누워있는 자세에서 변하지 않아 다른 행동과는 명확히 구분되어 100% 인식했으며 행동 없음의 경우, 약 먹기 데이터에서 분리하여 정답 값을 만들었기 때문에 두 클래스 간의 영상 유사성으로 인한 오분류로 다른 클래스 대비 성능이 떨어지는 모습을 보였다.

Table 5. Results of Action Recognition Model

Class	Recall	Precision	F1-score
None	0.77	0.5	0.61
Take Medicine	0.86	0.83	0.84
Remote Control	0.88	0.97	0.92
Fallen	1	1	1
Mean	0.88	0.82	0.85

3. Inference Result of Edge Device

돌봄 로봇의 엣지 디바이스인 VIM4에 표정 인식 모델과 행동 인식 모델을 이식해 추론 시간을 측정하였다. VIM4의 사양은 표 6과 같다.

Table 6. Specifications of Edge device Used for Hyodol

CPU	2.2GHz Quad core ARM Cortex-A73 2.0GHz Quad core Cortex-A53
GPU	ARM Mali-G52 MP8(8EE) GPU
RAM	8GB LPDDR4X
OS	Android 11

표 7은 VIM4 디바이스에서 각 모델에 대한 추론 시간 결과이다. ML Kit를 사용한 얼굴 검출 모델의 추론 속도는 20ms로 측정되었으며, 엣지 디바이스에서 구동되는 것을 목적으로 만들어진 SDK이기 때문에 실시간 처리에 적

합한 추론 속도를 보여주었다. 실험 결과는 표정 인식 모델과 행동 인식 모델은 각각 50ms, 350ms에 처리 속도를 보여주고 있다. 행동 인식 모델에서 영상의 특징을 추출하는 백본인 비전 트랜스포머의 처리 속도는 330ms이며 여덟 프레임 간의 시간 정보를 학습해 행동 분류기 역할을 하는 MLP의 처리 속도는 20ms이다. 모든 실험 결과를 통해 제안하는 모델들과 시스템이 현재 출시된 돌봄 로봇의 고도화에 적합하고 적용 가능하다는 점을 알 수 있다. VIM4에서의 구동을 확인하기 위한 용도로 유저 인터페이스를 제작하였고 그 예시는 그림 3에 나타내었다.

Table 7. Results of Edge Device Inference

Model		Inference Time(ms)
Face Detect		20
Facial Expression Recognition		50
Action Recognition	Backbone	330
	MLP	20

V. Conclusions

본 논문에서는 기존의 돌봄 로봇의 문제점인 접촉 방식을 벗어나 딥러닝을 이용한 표정 및 행동 인식을 통해 비접촉 방식으로 사용자와 상호작용할 수 있는 연구를 진행했다. 돌봄 로봇은 고령화가 진행되고 있는 한국에서 노인들을 보조해주며 부족한 인력을 해결할 수 있는 하나의 수단이다. 현재 한국에서 서비스되고 있는 돌봄 로봇은 주로 촉각 센서나 음성으로 정보를 수집해 도움을 주는 방법으로 개발되었다. 이에 따라 직접적인 접촉이 필요하였으나, 제안 시스템에서는 돌봄 로봇에 사용자의 표정과 행동을 인식하는 알고리즘을 추가하여 이를 극복하였다. 실험 결과를 통해 제안한 모델들의 성능과 속도가 돌봄서비스에 적합하다는 것을 확인하였다. 제안한 돌봄 로봇에 카메라를 달아 영상 서비스를 제공하는 시스템은 다양한 장점이 있다. 표정 인식으로 현재 사용자의 기분을 인지해 그에 맞는 서비스를 제공할 수 있고 행동 인식을 통해 사용자의 위급한 상황을 인식하고 대응할 수 있다. 추후 연구에서는 표정 및 행동 인식 기능을 고도화해 더 많은 표정과 행동을 인식하여, 더욱 다양한 사용자의 요구를 충족시켜 더 나은 돌봄서비스를 제공할 것을 목표로 한다.

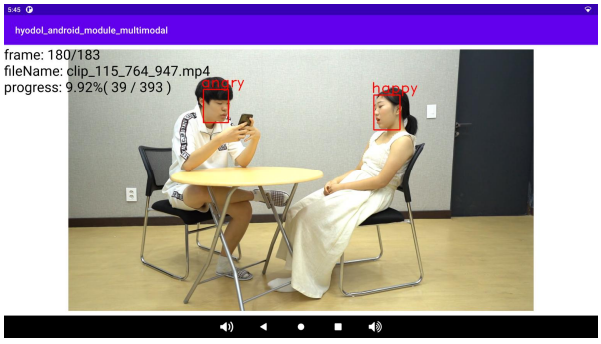

Fig. 3. Interface of Care Robot System in VIM4

ACKNOWLEDGEMENT

A research conducted with the funds from 「Local SW Service Commercialization Project('22~'23)」, Ministry of Science and ICT/Korea Information Society Agency.

This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program) (20009972, Support for commercialization of next generation semiconductor technology development) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

This work was supported by Police-Lab 2.0 Program(www.kipot.or.kr) funded by the Ministry of Science and ICT(MSIT, Korea) & Korean National Police Agency(KNPA, Korea) [Project : Video Analysis and Summary System Using Artificial Intelligence-Based Child Abuse Detection Technology) / Project Number : 230122M0101]

REFERENCES

[1] Statistics Korea, “Population Projection for Korea,” *https://www.kostat.go.kr/*

[2] J. Mišeikis et al., "Lio-A Personal Robot Assistant for Human-Robot Interaction and Care Applications," in IEEE Robotics and Automation Letters, vol. 5, no. 4, pp. 5339-5346, Oct 2020, DOI: 10.1109/LRA.2020.3007462

[3] Kim J-W, Choi Y-L, Jeong S-H, Han J. “A Care Robot with Ethical Sensing System for Older Adults at Home,” Sensors, 22(19):7515, Oct 2022. DOI: 10.3390/s22197515

[4] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M, "Blazepose: On-device real-time body pose tracking," arXiv preprint arXiv:2006.10204, Jun 2020, DOI: 10.48550/arXiv.2006.10204

[5] I. J. Goodfellow et al, "Challenges in representation learning: A report on three machine learning contests," in International Conference on Neural Information Processing. pp. 117–124, Nov 2013, DOI: 10.1007/978-3-642-42051-1_16

[6] Yousif Khaireddin and Zhuofa Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013," arXiv:2105.03588, May 2021, DOI: 10.48550/arXiv.2105.03588

[7] Y. Nan, J. Ju, Q. Hua, H. Zhang and B. Wang, "A-MobileNet: An approach of facial expression recognition," Alexandria Engineering Journal, vol 61(6), pp 4435-4444, Jun 2022, DOI: 10.1016/j.aej.2021.09.066

[8] Sangmin Kim, Seoung Bum Kim, "Abnormal Action Detection of Older Adults Using Graph Convolution Autoencoder," KOREAN MANAGEMENT SCIENCE REVIEW, 39(3), pp 29-43, Sep 2022, DOI: 10.7737/KMSR.2022.39.3.029

[9] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291-7299, Jul 2017, DOI: 10.48550/arXiv.1812.08008

[10] Zin TT, Htet Y, Akagi Y, Tamura H, Kondo K, Araki S, Chosa E. "Real-Time Action Recognition System for Elderly People Using Stereo Depth Camera," Sensors, vol. 21(17):5895, Sep 2021. DOI: 10.3390/s21175895

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv:2010.11929, Oct 2020. DOI: 10.48550/arXiv.2010.11929

[12] AI Hub, "Complex Video for Korean Emotion Recognition", *https://www.aihub.or.kr/*

[13] Tan, M., & Le, Q., "Efficientnet: Rethinking model scaling for convolutional neural networks," In International conference on machine learning, PMLR pp. 6105-6114, May 2019, DOI: 10.48550/arXiv.1905.11946

[14] A. V. Savchenko, L. V. Savchenko and I. Makarov, "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network," in IEEE Transactions on Affective Computing, vol. 13, no. 4, pp. 2132-2143, Jul 2022, DOI: 10.1109/TAFFC.2022.3188390

[15] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. "Vggface2: A dataset for recognising faces across pose and age," In 2018 13th IEEE international conference on automatic face & gesture recognition, pp. 67-74. Jun 2018, DOI: 10.1109/FG.2018.00020

[16] Jinhyeok Jang, Dohyung Kim, Cheonshu Park, Minsu Jang, Jaeyeon Lee, Jaehong Kim, "ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly," International Conference on Intelligent Robots and Systems (IROS) 2020, pp. 10990-10997. Feb 2020, DOI: 10.1109/IROS45743.2020.9341160

[17] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. "Learning transferable visual models from natural language supervision," In International conference on machine learning, pp. 8748-8763), PMLR, Jul 2021, DOI: 10.48550/arXiv.2103.00020

[18] Junsik Lee, In-Jin Yoo, Do-Hyung Park, "Implementation Strategy for the Elderly Care Solution Based on Usage Log Analysis: Focusing on the Case of Hyodol Product," Journal of Intelligence and Information Systems, vol. 25(3), pp. 117-140, Sep 2019, DOI: 10.13088/jiis.2019.25.3.117

[19] KHADIS, Khadis VIM4(KVIM4-B-001), *https://www.khadas.com/product-page/vim4.*

Authors

Hyun-sik Ham received the B.S. degree in Electrical and Electronic Engineering and M.S. degrees in Interdisciplinary Graduate Program for BIT Medical Convergence from Kangwon National University, South Korea, in 2020 and 2022, respectively. Mr. Ham joined ZIOVISION Co., Ltd. at Chuncheon in 2022 He currently works as a researcher at ZIOVISION.

Sae Jun Ko received the B.S. degree in Computer Engineering from Kangwon National University, Korea, in 2023. Mr. Ko joined ZIOVISION Co., Ltd. at Chuncheon in 2023 He currently works as a researcher at ZIOVISION.