카메라 및 관성 측정 장치 융합을 활용한 딥러닝 기반 수작업 조립 공정 작업자 동작 인식 방법론 개발

Development of Deep Learning-based Worker's Activities Recognition Methodology using Camera and Inertial Measurement Unit Fusion

양동혁 연구원[1], 신희진 연구원[1], 김도현 연구원[1], 강용신 연구실장[1]

[1]차세대융합기술연구원 산업지능연구실

{dhyang, gmlwls2407, jdh3643, yskang}@snu.ac.kr

초록

본 연구에서는 수작업 조립 공정 내에 작업자에게 방해를 주지 않으면서 동시에 효과적으로 동작을 인식하는 멀티모달 딥러닝 기반 작업자 동작 인식 방법론을 제안한다. 원거리 카메라와 관성 측정 장치가 내장된 스마트 워치를 사용해 영상 및 센서 데이터 형태로 작업자의 동작 데이터를 수집한 후, 특징 추출 기법을 적용해 작업자 동작 데이터로부터 특징을 추출했다. 유니모달 및 멀티모달 딥러닝 모델을 활용해 작업자 동작 인식을 수행한 결과 곱셈 규칙을 적용한 멀티모달 딥러닝 모델이 가장 우수한 성능을 보였고 본 연구에서 제안한 방법론이 실제 산업 현장에서도 적용 가능하다는 것을 보였다.

1. 서론

사람 행동 인식(Human Activity Recognition, 이하 HAR)이란 카메라, 센서 등 다양한 장치를 활용하여 사람의 행동 데이터를 수집한 후, 딥러닝(Deep Learning)과 같은 기법을 활용하여 사람의 행동을 인식하는 기술을 의미한다(Gupta *et al.*, 2002). 최근 카메라 및 센서 비용이 감소하고 컴퓨팅 기술이 발전하면서 HAR은 다양한 분야에서 활용되고 있으며(Li et al., 2016; Lu *et al.*, 2020), 특히 제조 분야의 경우 운영 효율성 향상(Chen *et al.*, 2020), 작업자 및 로봇 협업 촉진(Wang *et al.*, 2019), 운영자 지원(Tao *et al.*, 2020), 작업자 교육 지원(Patalas-Maliszewska *et al.*, 2021), 그리고 생산성 및 안전성 향상(Kobayashy *et al.*, 2019) 등 다양한 이유로 HAR 관련 연구가 활발히 진행중이다.

HAR 관련 연구에는 크게 영상 기반 HAR과 센서 기반 HAR 두 가지로 구분할 수 있다(Mekruksavanich and Jitpattanakul, 2021). 영상 기반 HAR을 수행하기 위한 방법 중 하나는, 카메라를 활용해 사람의 행동을 촬영한 영상을 연속된 이미지 데이터로 분할한 후 시·공간적 특징을 모두 반영할 수 있는 인공지능 모델을 활용해 사람의 행동을 인식하는 것이다(Khan *et al.*, 2022). 또 다른 방법은 수집한 영상 데이터에 광학 흐름(optical flow)과 같은 특징 추출(feature extraction) 기법을 활용해 영상 데이터로부터 시간적 특징이 반영된 이미지를 추출한 후, 딥러닝 모델 등을 활용해 사람의 행동을 인식하는 것이다(Xu *et al.*, 2021).

센서 기반 HAR을 수행하기 위한 대표적인 방법으로, 가속도(accelerometer) 및 자이로스코프(gyroscope) 센서를 포함하는 관성 측정 장치(Inertial Measurement Unit, 이하 IMU)가 내장된 웨어러블 센서를 활용해 행동 인식 대상자로부터 행동 데이터를 수집한 후, 시간적 특징을 반영할 수 있는 인공지능 기법 등을 활용하는 것이다(Wang and Liu, 2020). 또 다른 방법으로 수집한 센서 데이터에 슬라이딩 윈도우(sliding window) 기법을 활용해 전체 데이터를 여러 개의 구역으로 나누고, 각 구역을 대표(represent)할 수 있는 특징을 추출한 후 딥러닝과 같은 기법을 적용해 사람의 행동을 인식하는 방법이 있다(Bianchi *et al.*, 2019).

기존의 카메라 또는 IMU 단일 센서를 활용한 유니모달(uni-modal) HAR 방법을 통해 사람의 행동을 효과적으로 인식하는데 크게 기여를 했지만, 행동 인식 대상이 카메라 촬영 범위에서 벗어나면 영상 데이터 수집이 어렵고 센서에 내장된 배터리가 방전이 되면 센서 데이터가 수집이 안된다는 한계가 존재한다(Chen *et al.*,2017). 이러한 유니모달 HAR의 단점을 개선할 수 있는 방법으로 카메라 및 IMU 융합(fusion)을 활용한 멀티모달(multi-modal) HAR이 있으며(Elmadany *et al.*, 2018), 유니모달 HAR보다 더 우수한 성능을 보였다(Franco *et al.*, 2020; Ehatisham-Ul-Haq *et al.*, 2019; Elmadany *et al.*, 2018; Dawar *et al.*, 2018; Dawar and Kehtarnavaz, 2018).

하지만, 기존 멀티모달 HAR 연구는 실제 산업 현장이 아닌 실험실 수준의 연구가 주류를 이루고 있다는 한계가 존재한다 (Wei *et al.*, 2019; Imran and Raman, 2020; Ahmad and Khan, 2020; Zou *et*

본 연구는 산업통상자원부와 한국산업기술진흥원의 "국제공동개발사업"의 지원을 받아 수행된 연구결과임. (과제번호: P0022807)

al., 2019). 실험실 수준의 HAR에서는 인식 대상자를 수월하게 촬영하기 위해 근거리 카메라를 사용하고, 인식 대상자는 데이터 수집 전용 웨어러블 센서를 착용한다. 하지만, 실제 산업 현장에서는 작업자의 행동이나 동선이 근거리 카메라에 의해 방해를 받을 수 있고, 센서 데이터 수집을 위해 평소에 사용하지 않는 중·대형 웨어러블 센서를 착용함으로써 작업자에게 불편함을 줄 수 있다.

이와 같은 배경으로, 본 연구에서는 실제 수작업 조립 공정 내 작업자의 동작을 효과적으로 인식하기 위해, 작업자의 행동 및 동선에 방해를 주지 않는 원거리 카메라와 작업자에게 불편함을 주지 않는 스마트 워치를 활용한 딥러닝 기반 멀티모달 HAR 방법론을 제안한다. 본 연구는 다음과 같이 구성된다. 제2절에서는 기존 HAR 관련 연구를 소개한다. 제3절에서는 본 연구에서 제안하는 멀티모달 HAR 방법론을 단계별로 설명한다. 제4절에서는 제안 방법론을 적용해 작업자 동작 인식 성능을 도출하는 실험을 수행하고 그 결과를 분석함으로써 제안 방법론의 우수성을 검증한다. 마지막으로 제5절에서는 본 연구에 대한 결론과 향후 연구 방향을 제시한다.

2. 관련 연구

본 절에서는 영상 기반 유니모달 HAR, 센서 기반 유니모달 HAR, 그리고 멀티모달 HAR에 관한 연구 사례를 소개한다.

2.1 영상 기반 유니모달 HAR

영상 기반 HAR에서는 주로 카메라를 활용해 사람의 행동을 촬영한 후, 수집한 영상 데이터로부터 시·공간적 특징을 추출해 HAR을 수행한다. (Andrade-Ambriz *et al.*, 2022)은 영상 데이터에 3차원(3 Dimensional, 이하 3D) 합성곱 장단기 기억(Convolutional-Long Short Term Memory) 알고리즘을 적용해 HAR 모델을 개발했다. (Khan *et al.*, 2022)은 키넥트(kinect) 센서를 활용해 사람의 행동을 촬영한 영상 데이터로부터 스켈레톤(skeleton) 데이터를 추출한 후, 2차원 합성곱 신경망 장단기 메모리(2 Dimensional Convolutional Neural Network-Long Short Term Memory, 이하 2D CNN-LSTM)를 활용해 HAR을 수행했다. (Xu *et al.*, 2021)은 합성곱 신경망(Convolutional Neural Network, 이하 CNN) 기반 MotionNet 알고리즘을 활용해 광학 흐름 이미지를 추출한 후 2D CNN을 활용한 HAR 프레임워크를 소개했다. (Stergiou and Poppe, 2021)은 영상 데이터로부터 시간적 특징을 추출할 수 있는 다중 시간 합성곱(multi-temporal convoultion)을 활용한 HAR 모델을 제안했다.

2.2 센서 기반 유니모달 HAR

센서 기반 HAR에서는 주로 IMU를 사람에 부착한 후 수집한 센서 데이터를 활용해 HAR을 수행한다. (Ahmad and Khan, 2021)은 센서 데이터에서 발생하는 급격한 변화(sharp transition)를 효과적으로 반영하기 위해 센서 데이터를 이미지로 변환하여 HAR을 수행하는 방법을 제안했다. 이를 위해 IMU를 통해 수집한 센서 데이터에 4가지 이미지 변환 기법(Signal Images, GAF(Gramian Angular Field) Images, MTF(Markov Transition Field) Images, RP(Recurrence Plot) Images)을 적용해 이미지로 변환했고, 2D CNN 기반 ResNet-18 모델을 활용해 HAR을 수행했다. (Abdu-Aguye *et al.*, 2020)은 1D CNN 기반 HAR 모델에 적응형 풀링 레이어(Adaptive pooling layer)를 추가함으로써 하이퍼파라미터 튜닝(hyperparameter tuning) 없이 범용적으로 적용 가능한 HAR 모델을 제안했다. (Bianchi *et al.*, 2019) 역시 슬라이딩 윈도우 기법을 적용해 센서 데이터를 샘플링(sampling)한 후, 1D CNN 기반 HAR 모델을 적용했다. (Wang and Liu, 2020)은 수집한 센서 데이터에 첨도(kurtosis), 왜도(skewness)와 같은 시간 영역 특징(time-domain feature)과 푸리에 변환(Fourier Transfrom)과 빈도 영역 특징(frequency-domain feature)을 추출한 후 LSTM 네트워크를 통해 HAR을 수행하는 방법을 소개했다.

2.3 멀티모달 HAR

멀티모달 HAR은 두 종류 이상의 센서로부터 수집된 데이터 각각에 유니모달 HAR 모델을 구축한 후 점수 융합(score fusion), 피처 융합(feature fusion), 점증 상관 분석(Canonical Correlation Analysis)과 같은 융합 기법을 활용해 최종 인식 예측 결과를 산출한다. (Wei *et al.*, 2019)은 멀티모달 HAR에서 자주 쓰이는 오픈 데이터셋(UTD-MHAD)을 활용해 영상과 센서 데이터를 모두 활용하는 방법을 제안했다. 제안된 연구에서 영상 데이터의 경우 3D CNN 모델을 활용했고, 센서 데이터는 경우 2D 관성 이미지로 변환 후 2D CNN 모델을 활용했다. 그 후 점수 융합과 피처 융합 방법을 사용해 멀티모달 HAR을 수행했다. 이와 비슷하게 (Imran and Raman, 2020)에서도 UTD-MHAD을 활용했다. 영상 데이터의 경우 SDFDI(Stacked Dense Flow Difference Image) 기법을 적용해 영상 데이터를 광학 흐름 기반 이미지로 변환했고, 스켈레톤 데이터와 센서 데이터의 경우 노이즈(noise)에 강한 데이터셋 구축을 위해 데이터 증강(data augmentation) 기법을 활용했다. 그리고 SDFDI, 스켈레톤, 센서 데이터 각각에 2D CNN, GRU(Gated Recurrent Unit), 그리고 1D CNN 알고리즘을 적용해 유니모달 HAR 모델을 구축한 후 점수 융합, 피처 융합, 점증 상관 분석을 적용해 멀티모달 HAR을 수행했다. (Ahmad and Khan, 2020)은 점수 융합, 피처 융합과 같은 단일 수준(single level)의 융합보다 더 우수한 융합 방법을 제안했다. 제안된 연구에서 기존에 자주 쓰이는 UTD-MHAD와 더불어 Berkeley MHAD와 UTD-MHAD Kinect V2 데이터셋까지 총 3개의 오픈 데이터셋을 활용했다. 뎁스(depth) 데이터를 뎁스 이미지로 변환했고 센서 데이터를 신호(signal) 이미지로 변환한 후, 제안한 심층 다단계 멀티모달 융합 프레임워크(Deep Multi-level Multimodal Fusion Frame)를 활용해 두 단계에 걸쳐 융합을 진행하고 HAR을 수행했다. (Zou *et al.*, 2019)은 와이파이(WiFi) 기반 데이터 전송 기능을 수행하는 IoT 장치와 RGB 카메라를 모두 활용하는 WiVi(Wifi-Vision) 멀티모달 HAR 방법을 제안했다. 제안된 연구에서, 수집한 와이파이 데이터와 영상 데이터에 각각 3D CNN와 C3D(Convolutional 3D) 모델을 적용했고, 앙상블(ensemble) 기법과 점수 융합을 적용한 HAR 모델을 개발했다.

3. 제안 방법론

본 절에서는 카메라 및 IMU 센서 융합과 딥러닝 기법을 활용해 수작업 조립 공정 내 작업자의 동작을 인식하는 과정에 대해 설명하며, 전체 과정은 1) 데이터 수집, 2) 데이터 가공, 3) 데이터 특징 추출, 4) 동작 인식 모델 구축, 5) 동작 인식 모델 융합으로 구성되어 있다.

3.1. 데이터 수집

본 연구에서는 공기 정화 장치를 생산하는 공장 내 팬 필터 유닛(Fan Filter Unit, 이하 FFU) 수작업 조립 공정에서 작업하는 작업자로부터 동작 데이터를 수집했다. 작업자가 하나의 FFU 조립 작업을 완수하기 위해 8개의 동작을 수행했으며, 총 44분 13초 동안 작업을 18회 반복 수행했다. 작업자가 수행한 동작에 대한 상세한 설명은 아래 <표 1>과 같다.

<표 1> 작업자 수행 동작 상세 설명

번호	동작	소요 시간 (초)	
		평균	표준편차
1	Unpackaging	8.5	2.9
2	Loading	4.9	0.8
3	Picking Up-Bolt	4.7	1.1
4	Bolting-Downside	11.1	15.2
5	Picking Up-Panel	6.5	1.8
6	Bolting-Panel	27.3	15.1
7	Reversing	3.0	0.8
8	Bolting-Upside	29.2	5.1

영상 데이터를 수집하기 위해 작업자의 동작을 관찰할 수 있는 곳에 카메라를 설치해 15 프레임 속도(Frames Per Second, 이하 FPS)로 영상을 촬영했고, 센서 데이터를 수집하기 위해 작업자가 작업을 수행하는 오른쪽 손목에 IMU가 내장된 스마트 워치(갤럭시 워치4)를 부착해 약 1 헤르츠(Hz)의 샘플링 속도(sampling rate)로 가속도 및 자이로스코프 센서 데이터를 수집했다.

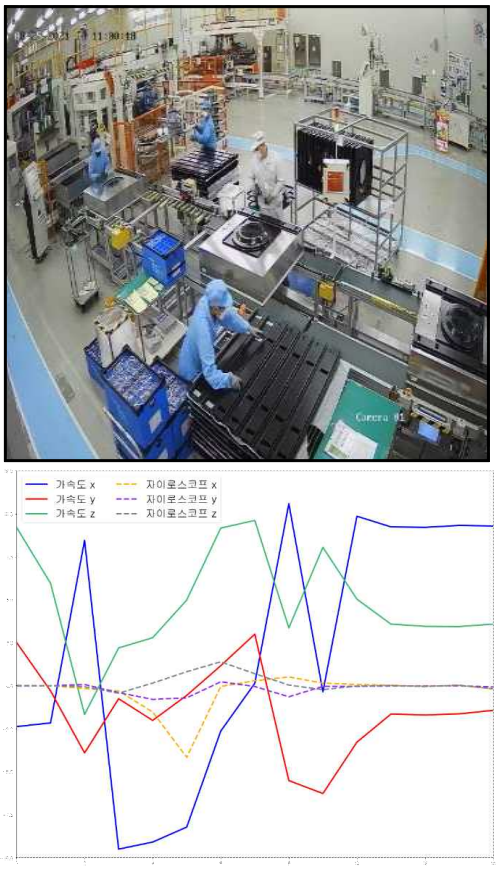
수집 결과, 총 39,795개의 이미지(1,440(높이) x 1,280(너비)) 데이터로 구성된 영상 데이터와 총 2,653개의 가속도 및 자이로스코프 x, y, z 3축 좌표 값으로 구성된 센서 데이터를 수집했다. 수집한 영상 및 센서 데이터 예시는 아래 <그림 1>과 같다.

3.2. 데이터 가공

작업자로부터 동작 데이터를 수집한 후 데이터 가공을 진행했다. 영상 데이터의 경우 작업자의 동작을 인식하는데 방해가 되는 배경 및 다른 작업자가 포함된 부분을 확인해 절삭(crop)했고, 그 결과는 아래 <그림 2>와 같다. 절삭 후 이미지 크기는 기존 1,440(높이) × 1,280(너비)에서 900(높이) × 600(너비)으로 축소되었다. 센서 데이터에서는 결측치(missing value)를 0으로 대치(imputation)했다.

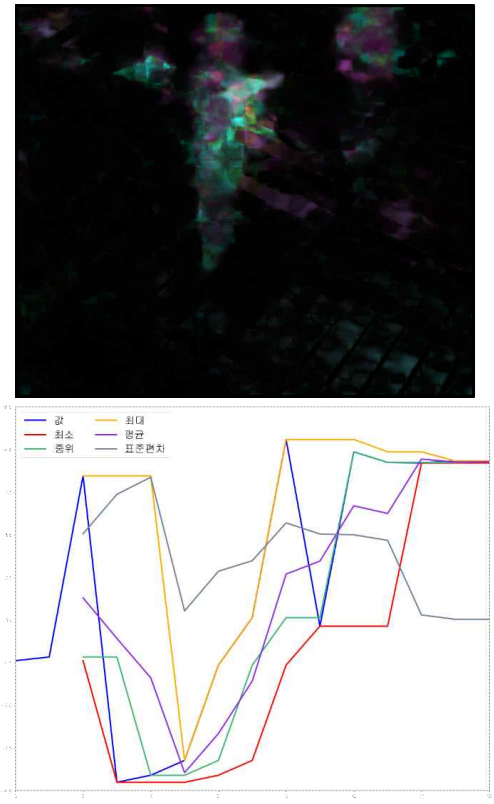
3.3 데이터 특징 추출

데이터 가공 후 데이터로부터 특징을 추출했다. 영상 데이터의 경우 광학 흐름(Farneback *et al.*, 2003) 기법을 적용해 연속된 30개의 이미지마다 특징이 추출된 하나의 이미지를 생성했고, 이를 수행하기 위해 *python calcOpticalFlowFarneback* 함수를 활용했다. 그 후, 특징이 추출된 전후의 이미지 데이터의 크기를 90(높이) × 60(너비)로 재조절(resize)했다. 센서 데이터의 경우 기간 이동 계산(rolling statistics)기법을 활용해 연속된 3개의 가속도 및 자이로스코프 3축 좌표 값마다 한 세트의 기술 통계량(최소, 중위, 최대, 평균, 표준편차)을 추출했고, 이를 계산하기 위해 *python rolling* 메소드를 활용했다. 특징이 추출된 데이터 예시는 아래 <그림 3>과 같다.



<그림 1> 수집 데이터 예시
(영상(상) / 센서(하))

<그림 2> 절삭 후 영상 데이터 예시



<그림 3> 수집 데이터 가공 결과 예시
(광학 흐름(상) / 기술 통계량(하))

3.4 동작 인식 모델 구축

데이터 특징 추출 후, 특징이 추출되기 전 데이터(이미지, 센서)와 특징이 추출된 데이터(광학 흐름, 기술 통계량)를 작업자 동작 인식 성능 평가에 활용했다. 활용 데이터에 대한 상세한 설명은 아래 <표 2>와 같다.

이미지, 광학 흐름, 센서, 기술 통계량 데이터를 학습하기 위해 각각 2D CNN-LSTM, 2D CNN, LSTM, 1D CNN 모델을 활용했고, 이를 위해 python tensorflow 라이브러리를 사용했다.

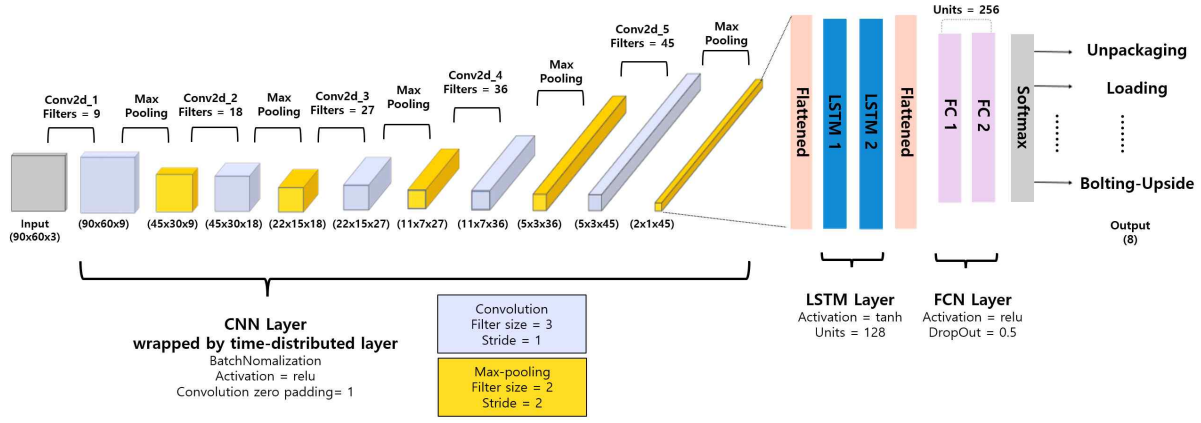
실험에 사용한 2D CNN-LSTM 모델은 아래 <그림 4>와 같이 크게 CNN, LSTM, Fully Connected Neural Network(완전 연결 신경망, 이하 FCN) 총 3개의 딥러닝 네트워크로 구성되어 있다. CNN 구조는 총 5개의 합성곱 층(convolution layer)으로 구성되어 있는데, 각각의 합성곱 층은 9, 18, 27, 36, 45개의 필터(filter)를 가지며, 필터 크기(filter size)는 3이고 스트라이드(stride)는 1이고, 배치 정규화(batch normalization) 기법을 적용했다. 인접한 합성곱 층 사이에는 최대 풀링(max pooling) 층이 존재하는데, 이 층의 필터 크기는 2이고 스트라이드가 2이다. 그리고 활성 함수(activation function)로 relu를 사용했다. LSTM 구조는 2개의 LSTM 층으로 구성되어 있으며, 각 LSTM 층에는 128개의 유닛(unit)이 존재하고 활성 함수로 tanh를 사용했다. FCN 구조는 2개의 FC 층으로 구성되어 있으며, 각 FC 층에는 256개의 유닛이 존재하고 활성 함수로 relu를 사용했다. 그리고 각 FC층에 0.5 비율만큼 드롭 아웃(drop out)을 사용했다. 실험에 사용한 2D CNN 모델과 1D CNN 모델은 2D CNN-LSTM 모델에서 CNN과 FCN의 결합과 같고, 실험에 사용한 LSTM 모델은 2D CNN-LSTM에서 LSTM과 FCN의 결합과 동일하다.

<표 2> 실험 데이터 상세 설명

데이터	데이터 크기	
	개수(개)	형태
이미지	39,795	90 × 60 × 3 (높이 × 너비 × 채널)
광학 흐름	1,325	90 × 60 × 3 (높이 × 너비 × 채널)
센서	2,653	6 (가속도/자이로스코프 센서 3축 좌표 수)
기술 통계량	1,325	30 (가속도/자이로스코프 센서 3축 좌표 수 × 기술 통계량 수)

3.5. 자세 인식 모델 융합

작업자 동작 인식 모델 구축 후, 서로 다른 두 모델을 융합하기 위해 점수 융합을 사용했다(Kittler et al., 1998). 점수 융합은 덧셈 규칙(sum rule), 곱셈 규칙(max rule), 최대 규칙(max rule)과 같은 규칙을 적용해 서로 다른 분류 모델에서 얻은 예측 확률 점수를 융합하는 방법이다. 점수 융합은 모델 추가 학습이 필요 없고 연산 시간이 짧다는 장점이 있어 다양한 멀티모달 HAR 연구에서 활용 중이다(Khaire *et al.*, 2018; Simonyan and Zisserman, 2014; Wang *et al.*, 2015).



<그림 4> 활용 2D CNN-LSTM 모델 구조

4. 실험 및 검증

본 절에서는 작업자 동작 인식 모델 성능을 평가하는 과정과 그 결과에 대해 설명한다.

4.1. 실험 계획

본 연구에서 다음과 같은 환경을 통해 실험을 진행했다. 실험에 사용한 컴퓨터의 CPU 사양(spec)은 Intel(R) Xeon(R) W-2245이고, GPU 사양은 NVIDIA Quadro RTX 4000이다. 작업자 동작 인식 모델을 학습하기 위해 사용한 하이퍼파라미터 중 옵티마이저(optimizer)로 Adam(Kingma and Ba, 2014)을 활용했고 , 학습률(learning rate)은 0.001, 베타 1(beta 1)은 0.9, 베타 2(beta 2)는 0.999를 적용했다. 에폭(epoch)은 30, 배치 크기(batch size)는 16을 사용했다. 2D CNN-LSTM 모델 학습을 위해 연속된 45개의 이미지를 하나의 입력 데이터로 활용했고, LSTM 모델 학습을 위해 연속된 3개의 센서 데이터를 하나의 입력 데이터로 활용했다. 총 20회 반복 실험을 진행했으며 각 실험마다 전체 데이터를 8대 2 비율로 학습 데이터(training set)와 테스트 데이터(test set)로 나눠 각각 모델 학습 및 성능 평가에 사용했다. 성능 평가 척도로 정확도(accuracy)를 활용했다.

4.2. 실험 결과

유니모달 딥러닝 HAR 모델을 활용한 작업자 동작 인식 실험 결과는 아래 <표 3>과 같다. 2D CNN-LSTM 모델의 정확도가 86.32%로 가장 우수한 인식 성능을 보였다. 그리고 2D CNN 모델(82.08%), 1D CNN 모델(69.92%), 그리고 LSTM(65.15%) 모델이 그 뒤를 따랐다. 주의할 만한 부분은 영상 기반 딥러닝 모델(2D CNN-LSTM, 2D CNN)의 정확도(80%대)가 센서 기반 딥러닝 모델(LSTM, 1D CNN)의 정확도(60%대)보다 높다는 것이다. 이를 통해 수작업 조립 공정 내 작업 중인 작업자의 동작을 인식하고자 할 때, 센서 기반 HAR 모델보다 영상 기반 HAR 모델을 활용하는 것이 더 낫다는 것을 알 수 있다.

<표 3> 유니모달 딥러닝 HAR 모델 기반 작업자 동작 인식 성능

데이터	딥러닝 알고리즘	정확도(%)
이미지	2D CNN-LSTM	86.32
광학 흐름	2D CNN	82.08
센서	LSTM	65.15
기술 통계량	1D CNN	69.92

멀티모달 딥러닝 HAR 모델을 활용한 작업자 동작 인식 실험 결과는 아래 <표 4>와 같다. 2D CNN-LSTM 모델의 결과와 1D CNN 모델동작 예측 결과를 곱셈 규칙으로 융합했을 때 정확도 93%로 가장 우수한 성능을 보였다. 그리고 이미2D CNN-LSTM 모델의 결과와 LSTM 모델의 결과를 곱셈 규칙으로 융합했을 때의 인식 정확도가 92.24%로 그 뒤를 따랐다. 주의할 만한 부분은 덧셈 또는 곱셈 규칙을 활용한 멀티모달 HAR 모델의 경우, 대부분 유니모달 HAR 모델보다 우수한 성능을 보였다는 것이다. 이를 통해 수작업 조립 공정 내에서 작업자의 동작을 인식하고자 할 때, 멀티모달 HAR 모델이 유니모달 HAR 모델보다 더 유효하다는 것을 알 수 있다.

<표 4> 멀티모달 딥러닝 HAR 모델 기반 작업자 동작 인식 성능

데이터	딥러닝 알고리즘	융합 규칙	정확도(%)
이미지 + 센서	2D CNN-LSTM + LSTM	덧셈	91.20
		곱셈	92.24
		최대	80.29
이미지 + 기술 통계량	2D CNN-LSTM + 1D CNN	덧셈	91.72
		곱셈	93.00
		최대	80.38
광학 흐름 + 센서	2D CNN + LSTM	덧셈	86.04
		곱셈	86.93
		최대	77.25
광학 흐름 + 기술 통계량	2D CNN + LSTM	덧셈	86.94
		곱셈	87.57
		최대	77.81

5. 결론

본 연구에서는 수작업 조립 공정 내에서 작업하는 작업자의 동작 데이터를 수집하고 작업자 동작 인식 성능을 도출하는 멀티모달 HAR 방법론을 제안했다. 이를 위해 원거리 카메라와 IMU가 내장된 스마트 워치를 사용해 작업자 동작 데이터를 수집했고, 광학 흐름과 기간 이동 계산 기법을 활용해 특징을 추출했다. 그 후 유니모달 딥러닝 모델을 활용해 작업자 동작 인식을 수행했고, 점수 융합 기반 멀티모달 작업자 동작 인식을 진행했다. 실험 결과 곱셈 규칙을 적용한 멀티모달 HAR 모델의 경우 가장 우수한 동작 인식 성능을 보였다.

본 연구에서 제안하는 방법론을 활용하면 작업자에게 불편함을 주지 않으면서도 작업자의 동작을 효과적으로 인식할 수 있으며, 작업자 안전 및 동작 효율성 향상, 최적 작업자 배치와 같은 추가 효과를 기대할 수 있다.

향후, 본 연구에서 제안하는 방법론을 개선할 수 있는 연구로 딥러닝 모델 구축 시 필요한 하이퍼파라미터 튜닝에 소요되는 시간을 줄일 수 있는 메타 휴리스틱(meta heuristics) 기반 딥러닝 하이퍼파라미터 최적화가 있다. 또한, 노이즈에 강건한(robust) 데이터셋 구축을 위해 방향 기울기 히스토그램(histogram of oriented gradients)과 같은 데이터 증강 기법을 적용한 HAR 프레임워크 개발이 있다.

참고문헌

Gupta, N., Gupta, S. K., Pathak, R. K., Jain, V., Rashidi, P., & Suri, J. S. (2022). Human activity recognition in artificial intelligence framework: A narrative review. Artificial intelligence review, 55(6), 4755-4808.

Wei, H., Jafari, R., & Kehtarnavaz, N. (2019). Fusion of video and inertial sensing for deep learning-based human action recognition. Sensors, 19(17), 3680.

Li, X., Zhang, Y., Li, M., Marsic, I., Yang, J., & Burd, R.

S. (2016, October). Deep neural network for RFID-based activity recognition. In Proceedings of the Eighth Wireless of the Students, by the Students, and for the Students Workshop (pp. 24-26).

Lu, M., Hu, Y., & Lu, X. (2020). Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. Applied Intelligence, 50, 1100-1111.

Chen, C., Wang, T., Li, D., & Hong, J. (2020). Repetitive assembly action recognition based on object detection and pose estimation. Journal of Manufacturing Systems, 55, 325-333.

Wang, L., Gao, R., V?ncza, J., Kr?ger, J., Wang, X. V., Makris, S., & Chryssolouris, G. (2019). Symbiotic human-robot collaborative assembly. CIRP annals, 68(2), 701-726.

Tao, W., Al-Amin, M., Chen, H., Leu, M. C., Yin, Z., & Qin, R. (2020). Real-time assembly operation recognition with fog computing and transfer learning for human-centered intelligent manufacturing. Procedia Manufacturing, 48, 926-931.

Patalas-Maliszewska, J., Halikowski, D., & Dama?evi?ius, R. (2021). An automated recognition of work activity in industrial manufacturing using convolutional neural networks. Electronics, 10(23), 2946.

Kobayashi, T., Aoki, Y., Shimizu, S., Kusano, K., & Okumura, S. (2019, November). Fine-grained action recognition in assembly work scenes by drawing attention to the hands. In 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 440-446). IEEE.

Mekruksavanich, S., & Jitpattanakul, A. (2021). Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models. Electronics, 10(3), 308.

Khan, I. U., Afzal, S., & Lee, J. W. (2022). Human activity recognition via hybrid deep learning-based model. Sensors, 22(1), 323.

Xu, J., Song, R., Wei, H., Guo, J., Zhou, Y., & Huang, X. (2021). A fast human action recognition network based on spatio-temporal features. Neurocomputing, 441, 350-358.

Wang, L., & Liu, R. (2020). Human activity recognition based on wearable sensor using hierarchical deep LSTM networks. Circuits, Systems, and Signal Processing, 39, 837-856.

Bianchi, V., Bassoli, M., Lombardo, G., Fornacciari, P., Mordonini, M., & De Munari, I. (2019). IoT wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment. IEEE Internet of Things Journal, 6(5), 8553-8562.

Chen, C., Jafari, R., & Kehtarnavaz, N. (2017). A survey of depth and inertial sensor fusion for human action recognition. Multimedia Tools and Applications, 76, 4405-4425.

Elmadany, N. E. D., He, Y., & Guan, L. (2018). Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis. IEEE Transactions on Multimedia, 21(5), 1317-1331.

Andrade-Ambriz, Y. A., Ledesma, S., Ibarra-Manzano, M. A., Oros-Flores, M. I., & Almanza-Ojeda, D. L. (2022). Human activity recognition using temporal convolutional neural network architecture. Expert Systems with Applications, 191, 116287.

Ahmad, Z., & Khan, N. (2021). Inertial sensor data to image encoding for human action recognition. IEEE Sensors Journal, 21(9), 10978-10988.

Abdu-Aguye, M. G., Gomaa, W., Makihara, Y., & Yagi, Y. (2020, July). Adaptive Pooling Is All You Need: An Empirical Study on Hyperparameter-insensitive Human Action Recognition Using Wearable Sensors. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE.

Imran, J., & Raman, B. (2020). Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition. Journal of Ambient Intelligence and Humanized Computing, 11, 189-208.

Ahmad, Z., & Khan, N. (2019). Human action recognition using deep multilevel multimodal (${M}^{2} $) fusion of depth and inertial sensors. IEEE Sensors Journal, 20(3), 1445-1455.

Zou, H., Yang, J., Prasanna Das, H., Liu, H., Zhou, Y., & Spanos, C. J. (2019). WiFi and vision multimodal learning for accurate and robust device-free human activity recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 0-0).

Farneb?ck, G. (2003). Two-frame motion estimation based on polynomial expansion. In Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29-July 2, 2003 Proceedings 13 (pp. 363-370). Springer Berlin Heidelberg.

Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. IEEE transactions on pattern analysis and machine intelligence, 20(3), 226-239.

Khaire, P., Kumar, P., & Imran, J. (2018). Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recognition Letters, 115, 107-116.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27.

Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., & Ogunbona, P. O. (2015). Action recognition from depth maps using deep convolutional neural networks. IEEE Transactions on Human-Machine Systems, 46(4), 498-509.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Franco, A., Magnani, A., & Maio, D. (2020). A multimodal approach for human activity recognition based on skeleton and RGB data. Pattern Recognition Letters, 131, 293-299.

Ehatisham-Ul-Haq, M., Javed, A., Azam, M. A., Malik, H. M., Irtaza, A., Lee, I. H., & Mahmood, M. T. (2019). Robust human activity recognition using multimodal feature-level fusion. IEEE Access, 7, 60736-60751.

Elmadany, N. E. D., He, Y., & Guan, L. (2018). Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis. IEEE Transactions on Multimedia, 21(5), 1317-1331.

Dawar, N., Ostadabbas, S., & Kehtarnavaz, N. (2018). Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. IEEE Sensors

Letters, 3(1), 1-4.
Dawar, N., & Kehtarnavaz, N. (2018, June). A convolutional neural network-based sensor fusion system for monitoring transition movements in healthcare applications. In 2018 IEEE 14th International Conference on Control and Automation (ICCA) (pp. 482-485). IEEE.
Stergiou, A., & Poppe, R. (2021, July). Multi-temporal convolutions for human action recognition in videos. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-9). IEEE.