종합설계2 프로젝트 제안서

객체 탐지(Object detection)와 인간 행동인지(Human action recognition)를 이용한 구조요청행동 인식 모델 구현

|  |  |  |
| --- | --- | --- |
| 지도교수 | 김기천 교수님 | |
| 팀 명 | 공 조 | |
| 팀 원 | 201212519 | 김선우 |
| 201615007 | 문기태 |
| 201411684 | 박재헌 |

**목 차**

1. 개요 및 목표

0

* 1. 프로젝트 배경
  2. 프로젝트 목표

1. 관련 기술 동향 분석
   1. 실시간 객체 인식 기술 (Real-time Object Detection)
   2. 실시간 자세예측 기술 (Real-time Pose Estimation)

2.3 실시간 인간행동 인지기술(Real-time Human Action Recognition)

1. 세부 개발내용
   1. 구조설계도
   2. 기술개념도
2. 진행일정
3. 팀원 별 역할
4. 참고문헌

1. 개요 및 목표

1.1 프로젝트 배경

**양팔을 들어 양 옆으로 흔드는(Waving arms) 행동**은 국제적으로 도움을 요청할 때 많이 사용하는 신호다. 이 행동은 다른 행동과 구별되는 특성이 있는데, **팔을 든 상태(Status)** 내에서 이루어지는 **반복적인(Repeatly) 행동**이라는 것이다.



그런데 기존 **객체탐지(Object detection)모델**에서는 정지된 프레임 내에서 **분류(Classification)** 문제를 다루기 때문에 프레임 간 **관계정보(Relation information)**가 필요한 움직임을 판단할 수 없다.

따라서, 연속된 프레임 간의 관계정보를 활용하는 **인간행동인식(Human action recognition) 모델**의 활용이 필요하다. 그런데, 이 모델은 현재 기술 수준의 **다중 객체인식(Multi Object recognition)에서 성능에 큰 문제**가 있다. 따라서 양손을 들어서 흔드는 동작에는 **행동 시작점(Action start point)**이 있다는 특성에 착안하여, 모델을 **손을 흔드는 동작이라는 주제로 구체화하여, 다중 객체 인식문제에 대해 기존 모델의 성능을 향상**시키고자 한다.

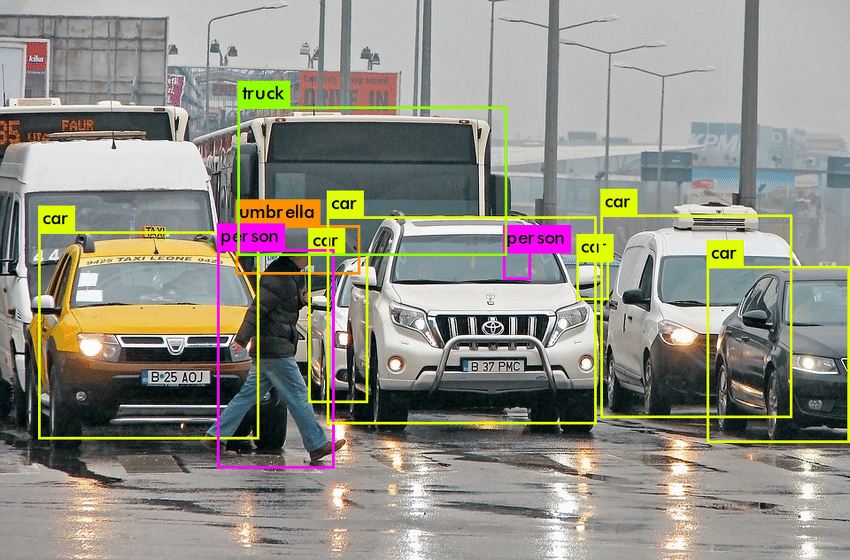
1.2 프로젝트 목표

현재, **인간 행동 인식 모델**은 객체 탐지 → 스켈레톤 추출 → 행동 인식의 **3step 구조**를 가지고 있다. 프로젝트의 모델을 구현하기 위해 먼저 객체탐지 부분에서 프레임 내 **손을 들고 있는 사람에 대해서만** **다음 스텝으로** 넘어가는 **모델을 구현**하고 성능비교를 하여, 손을 흔드는 특정 행동에 대한 **다중객체 상황에서 행동인식 모델의 성능을 향상**하고, **기존 모델과 성능을 비교**를 할 것이다. 성능 향상의 수준이 매우 뛰어나다면, **실시간(real time) 활용까지** 해볼 계획이다.

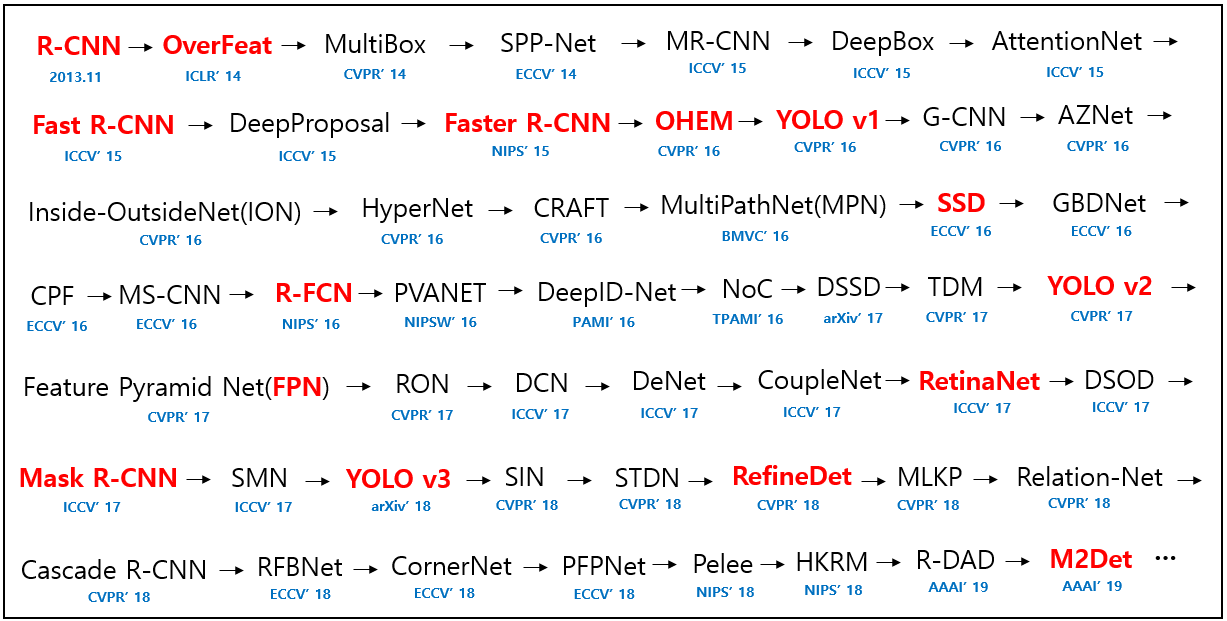
2. 관련 기술 동향 분석

2.1 실시간 객체 탐지(Realtime Object Detection)

**이미지나 동영상 프레임에서 사람, 동물 등 의미 있는 개체(Object)의 종류와 위치를 찾는 기술이다.**



**-기술 발전 동향(논문 발표 시기 기준)**



위 사진의 빨간색 글씨 부분은 자주 사용되는 모델들이다. 객체 탐지 모델의 성능은 정확성과 처리속도에 좌우되는데, 실시간 수준의 처리 속도를 가진 모델로 Fast R-CNN, SSD, Yolo-v3를 꼽을 수 있다. 이중에 처리 속도 대비 상당 수준의 정확도를 가진 Yolo-v3 모델을 본 프로젝트에서 활용할 것이다.

2.2 2D 실시간 포즈 예측 기술(2D real-time pose estimation)

**이미지나 영상 프레임에서 인간의 몸과 손, 얼굴, 관절 등의 키포인트(Key point)를 찾아서 이미지에서 객체를 식별하는 것(Object Detection)을 넘어 개체의 자세를 추측하는 기술이다. (Object Detection+ Pose Estimation)**

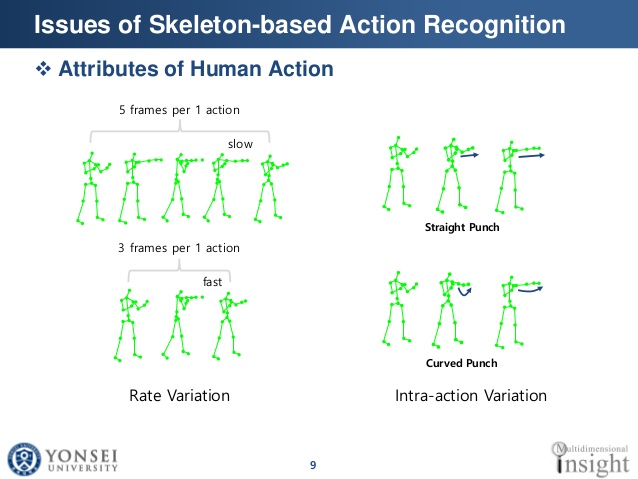


**- 기술 발전 동향**

기존에는 3d 카메라 또는 키넥트 센서를 이용해서 depth 데이터를 획득하였으나 최근 딥러닝 분야의 발전으로 일반 카메라로부터 예측된 데이터를 획득하는 것이 가능하다. Pose estimation은 2d 데이터를 사용하는 방법과 3d 데이터를 사용하는 방법으로 나뉜다. 2d 데이터를 사용하면 처리속도는 빨라지나 정확도가 낮고, 3d 데이터를 사용하면 정확도가 높은 대신 처리량이 많아져 속도가 늦어진다. 우리는 이번 프로젝트에서 depth정보가 없는 2D 비디오를 다루기 때문에, 이 분야에서 독보적으로 많이 사용되는 OPENPOSE 모델을 활용할 것이다.

2.3 인간 행동 예측 기술(Human action recognition)

**이미지나 동영상 프레임에서 일정 시간 단위로 식별된 개체들의 변화를 파악하고 이를 통해 어떤 행동인지에 대해 추측하는 기술(frame per action) 이다.**

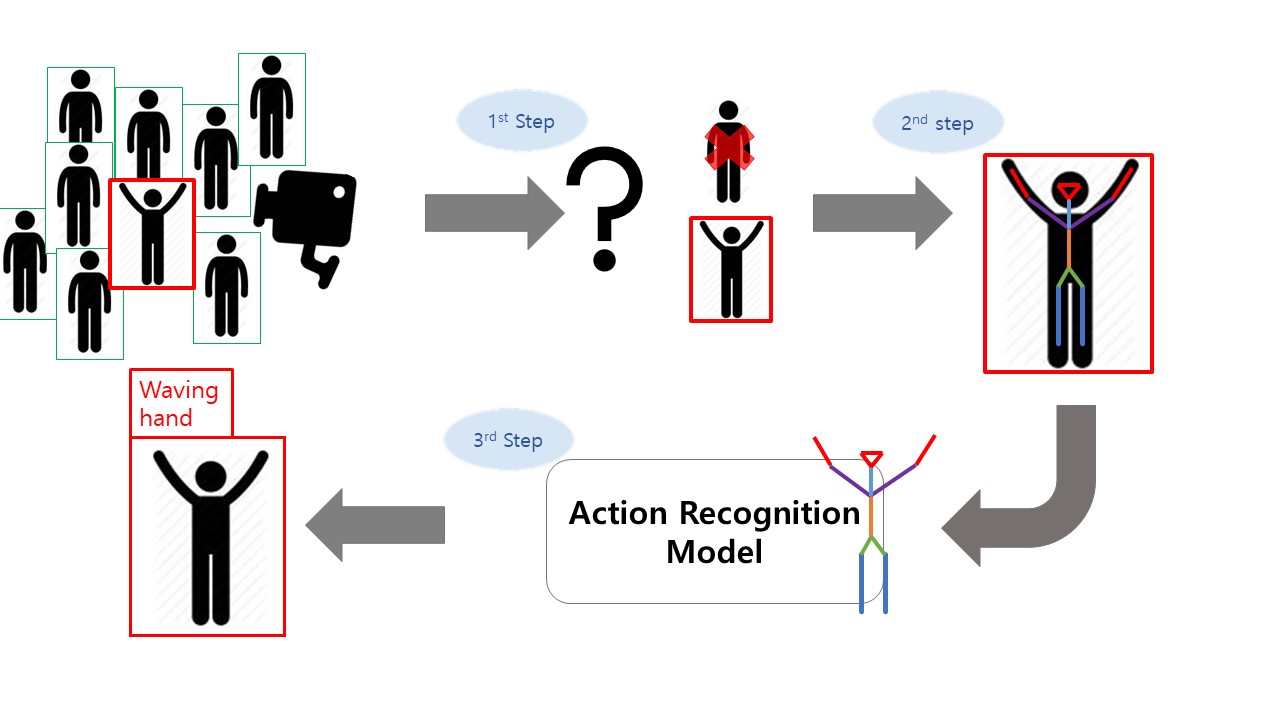


**-기술 발전 동향**

오늘날 행동인식 중에서는 손 동작 인식 기술이 따로 있을 만큼 손 동작을 인지하고 처리하는 기술의 연구가 활발히 일어나고 있다. 손은 사용자가 쉽게 사용할 수 있으며 자유도가 높아 다양한 표현을 할 수 있는 매개체이다. 이러한 손을 추적하며 동작을 분석하기 위해 과거에는 컬러센서를 통해 촬영된 영상에서 손을 찾아 분할하고 분할된 영상에서 손의 자세를 인식하는 등의 접근을 해 왔다. 이러한 컬러영상을 활용하던 방법들을 너머, 최근에는 거리영상을 얻을 수 있는 장비들이 상용화 됨에 따라 거리 정보를 이용하여 사용자 손을 분석하고 동작을 추정하는 연구들이 활발하게 진행되고 있다.

3. 세부 개발 내용

3.1 구조설계도



1st Step – **Yolo v3 모듈**을 이용한 인간 탐지와 손을 들고 있는 사람과 안 들고 있는 사람의 **구분**(Yolov3 module object Detection and Classification)

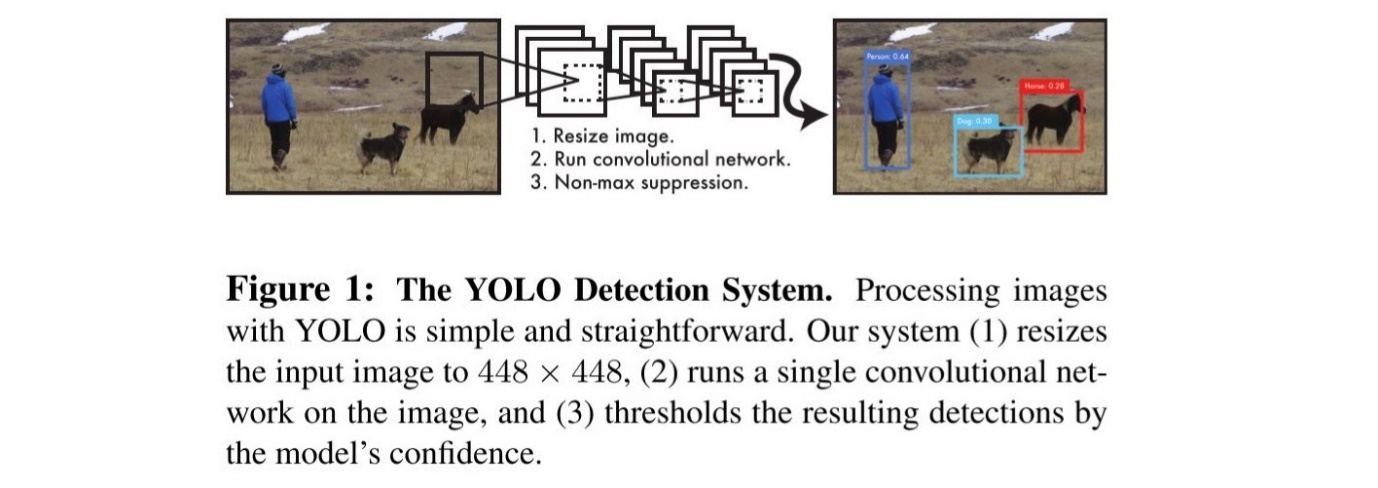
2nd Step – **Open pose 모듈**을 이용하여 손을 들고 있는 사람에게서만 **스켈레톤 추출.** (Open pose module pose estimation)

3rd Step – **LSTM 모델**을 이용한 인간 행동 인식 모듈을 이용하여 손을 든 포즈에 한하여 프레임간 릴레이션 분석을 통해 손을 흔드는지 여부를 **판단**. (human action recognition)

3.2 기술 개념도

1st Step) Yolo v3 객체 인식 및 분류 모듈

YOLO(You Only Look Once)는 이미지 내의 bounding box와 class probability를 single regression problem으로 간주하여, 이미지를 한 번 보는 것으로 object의 종류와 위치를 추측한다.



우리 프로젝트에서Yolov3 모델의 역할은 손을 흔드는 행동에 대한 시작점을 찾아서 2Step으로 넘어가는 **스위치(Switch)의 역할**이다. 앞에서 말했듯이 손을 흔드는 행동은 같은 상태(손을 들고 있음)내에서 같은 동작을 반복(Cyclic)하는 행위이다. 같은 종류의 행동으로 걷기(Walking)가 있다.

2nd Step) Openpose Module

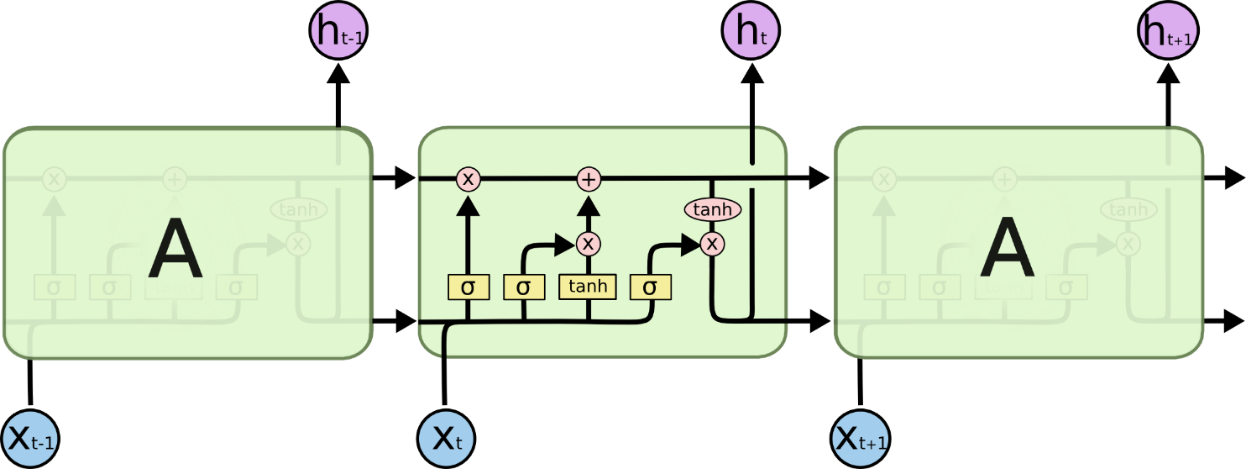
CPM을 Bottom-up 형태로 디자인해 효과적으로 개선한 모델이다. 사진 또는 영상에서 한 사람의 동작뿐만 아니라 여러 사람의 동작에서 몸, 손, 얼굴, 얼굴의 형태 등의 keypoint를 추출할 수 있다. OpenPose로부터 출력된 스켈레톤 데이터는 pose estimation 모델 학습에 사용될 수 있다.



오픈 포즈는 단 한사람의 포즈를 예측하는 데에도 **상당한 수준의 하드웨어가 필요**하다. 조사한 바에 따르면, 최소 요구 사항은 코어 intel i7 8thgen, 그래픽 카드 GTX 1080, 램 32GB다. 심지어, **객체가 다중으로 늘어나면 슈퍼 컴퓨터 정도의 사양이 필요하다.** 결론적으로, PC에서 돌리는 것이 불가능하다는 것이다. **여기에 우리 프로젝트의 의의가 있다.** 오픈 포즈 이전에 탐지 레이어를 추가 하여 **선별적 포즈 예측**을 함으로서, 다중 객체 상황에서 하드웨어의 부담을 줄여 전체 모델의 성능을 향상시킬 수 있는 것이다.

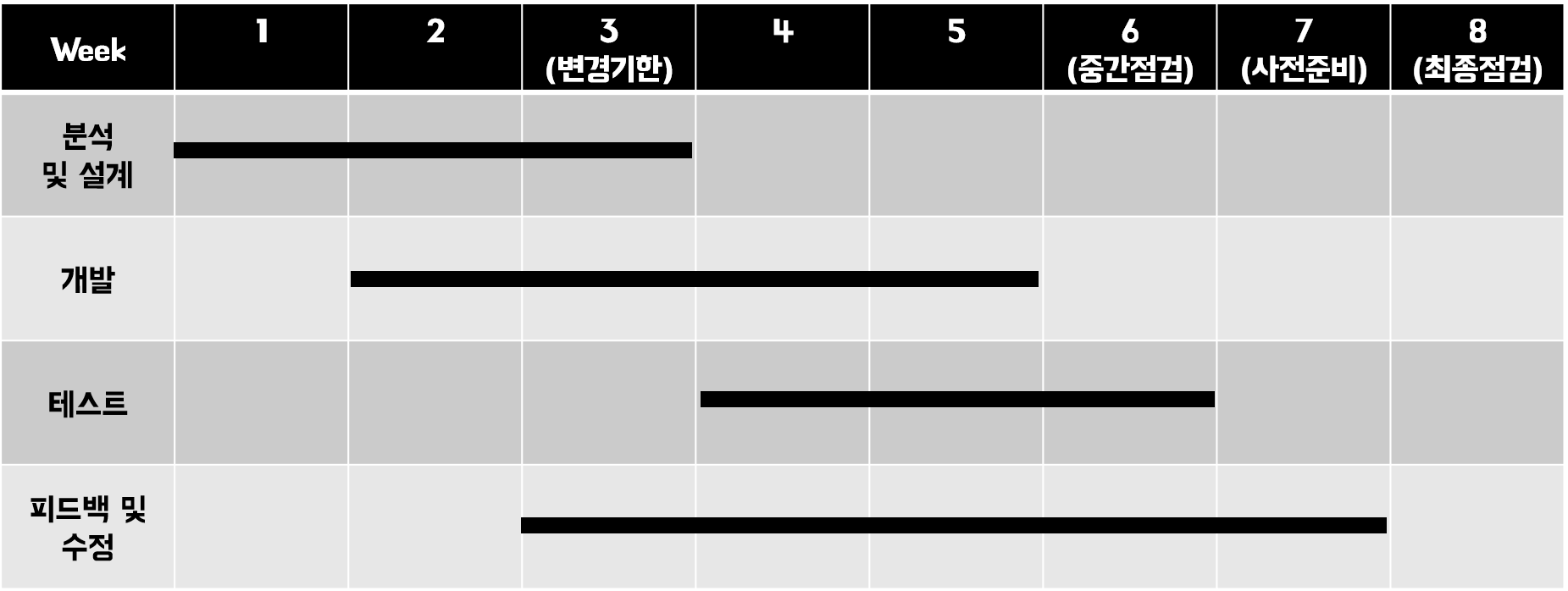
3rd Step) LSTM Action recognition Module

기존 RNN은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 역전파시 gradient가 점차 줄어 학습능력이 크게 저하되는 현상이 발생하는데 이를 해결하기 위한 모델이 바로 LSTM 모델이다. LSTM은 RNN의 Hidden state에 cell-state를 추가한 구조이다. cell state는 일종의 컨베이어 벨트 역할을 하게 되는데 이로 인해 state가 꽤 오래 경과하더라도 gradient가 비교적 전파가 잘 된다.



이를 Human Action Recognition에 적용하기 위해서는 먼저 가속도계 센서에서 수집 한 데이터를 사용한다. 이 모델은 가속도계 및 자이로 스코프 데이터의 각 축과 같은 여러 병렬 입력 데이터 시퀀스를 지원할 수 있기 때문에 일련의 관측치에서 피쳐를 추출하고 내부 피쳐를 다른 활동 유형에 매핑하는 방법을 학습한다.

4. 진행일정



Step1. 분석 및 설계: OOA Step을 통한 Requirement 분석 및 프로젝트 설계

Step2. 개발: 프로젝트 개발

Step3. 테스트: 개발된 output에 대한 Test Case기반의 테스트

Step4. 피드백 및 수정: 여러 Estimation을 거쳐 지속적인 수정보완 예정

5. 팀원 별 역할

팀원1 김선우: Object detection모델 담당, 전체적인 시스템 플로우 관리, 팀장

팀원2 박재헌: Pose estimation 모델 담당, Opencv활용 영상처리, 총무

팀원3 문기태: Human action recognition 모델 담당, Gui 및 기타업무

6. 참고문헌

[1] <https://www.themtank.org/a-year-in-computer-vision>

[2] Zhe Cao et al, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR, 2016

[3] Riza Alp Guller et al, DensePose: Dense Human Pose Estimation In The Wild, arXiv, 2018

[4] Shih-En Wei et al, Convolutional Pose Machines, 2016