Human Activity Recognition in Public Places

W251 Team Human Activity

Sean Campos
Sam Shih
Praveen Reddy Kasireddy
Fengyao Luo

Contents

- Background Introduction
- Data Pipeline
  - ML at Edge
  - Body Pose Detection
  - Body Framing
  - Action Recognition
- Results
- Simulation (Demo)
- Challenge and Next Steps

Use Case / Problem Statement

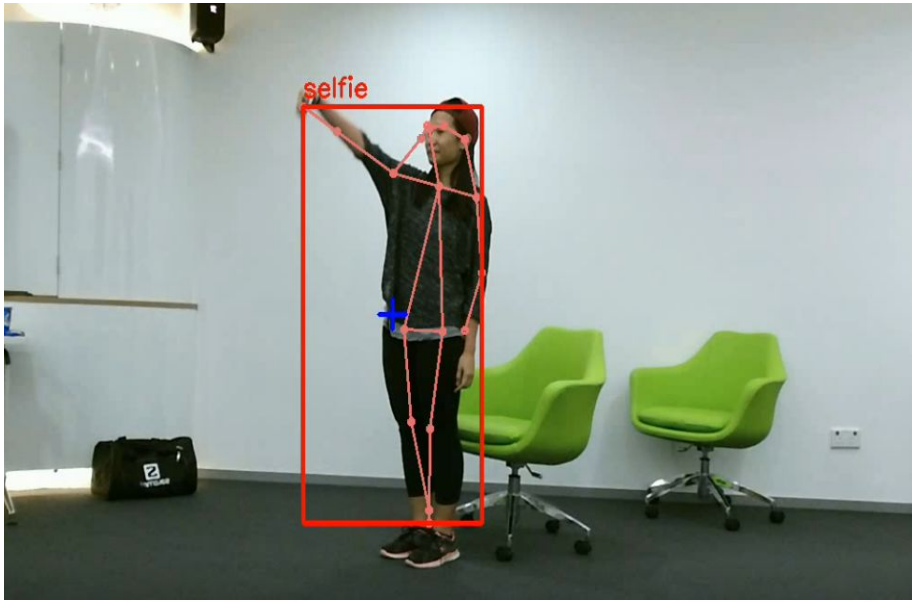
Problem: Public resource groups like parks & recreation would like to know how the public places are being used?

Traditional Solution: Engage people to physically scout the places and gather the data.

- Costly (Surveyor)
- Safety issues (Covid, Privacy)

Proposed Solution: ML and edge devices with video cameras.

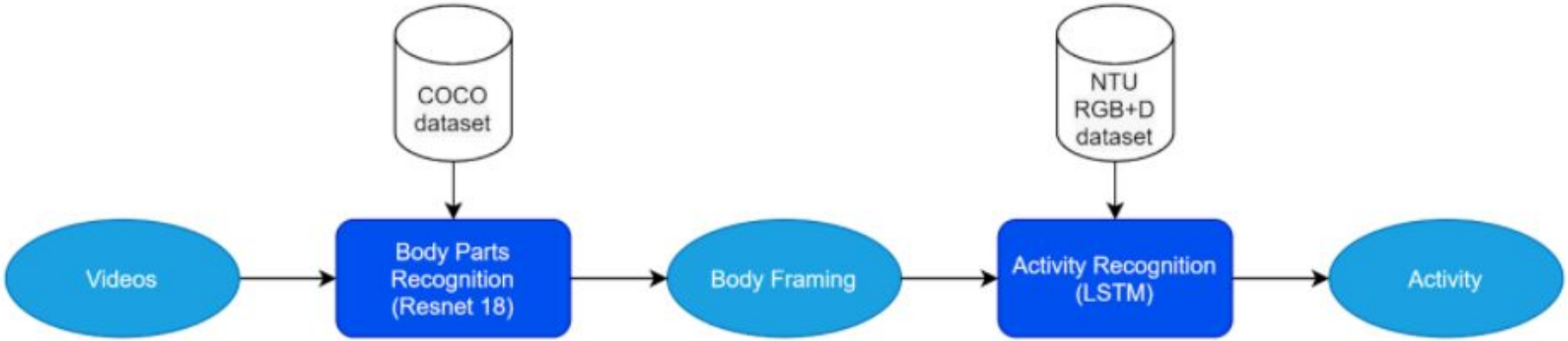
- Less cost & Scalable
- Safe and can manage better privacy (Superintendents)



Data Pipeline

Phase 1: Body Part Detection (Resnet18)

Phase 2: Action Recognition (LSTM)



Machine Learning at Edge

Camera:

- Stream video to the inference engine on the Edge Device

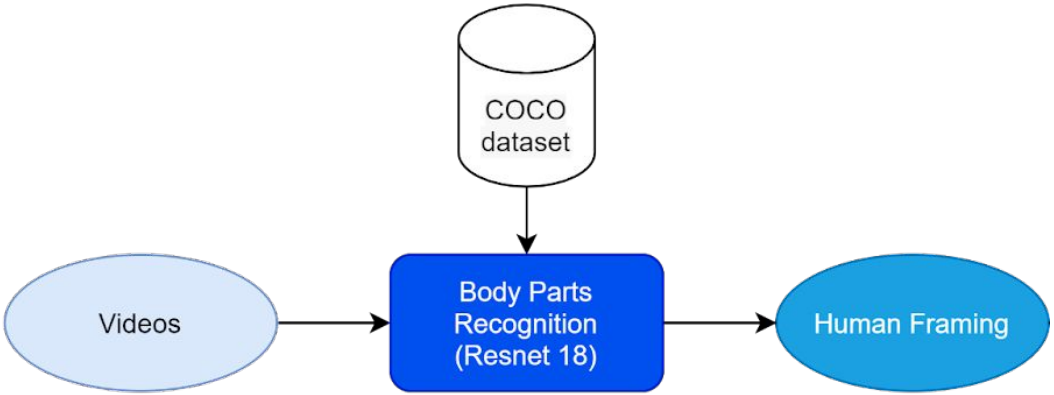
Edge Device(Jetson):

- Extract information about people and their body parts
- Perform object tracking
- Detect activity
- Record the metrics

Body Part detection - Solution

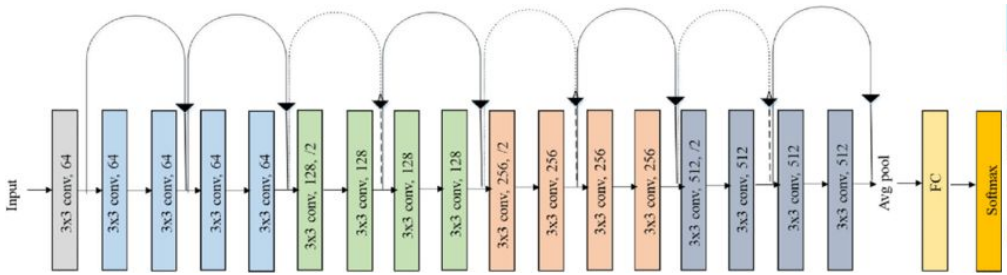
Transfer Learning:

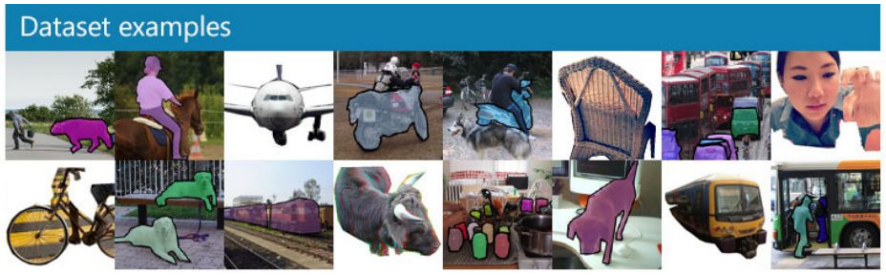
- Used pre-trained model (Resnet18)
- Fine tune with COCO dataset
- Identify body parts
- Object tracking



Body Part detection - Pretrained Model

ResNet18

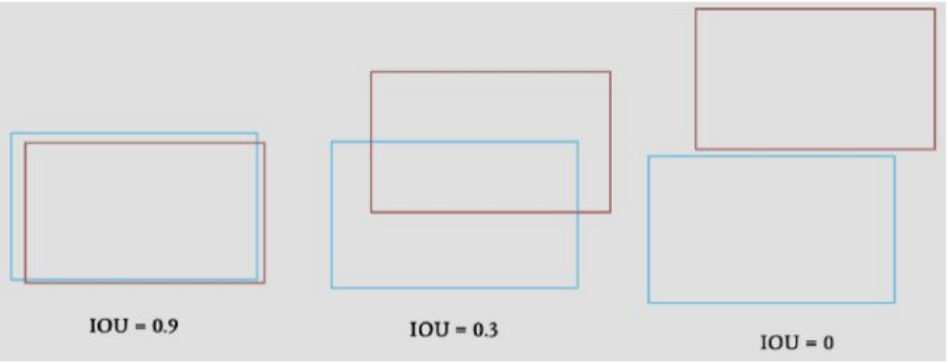


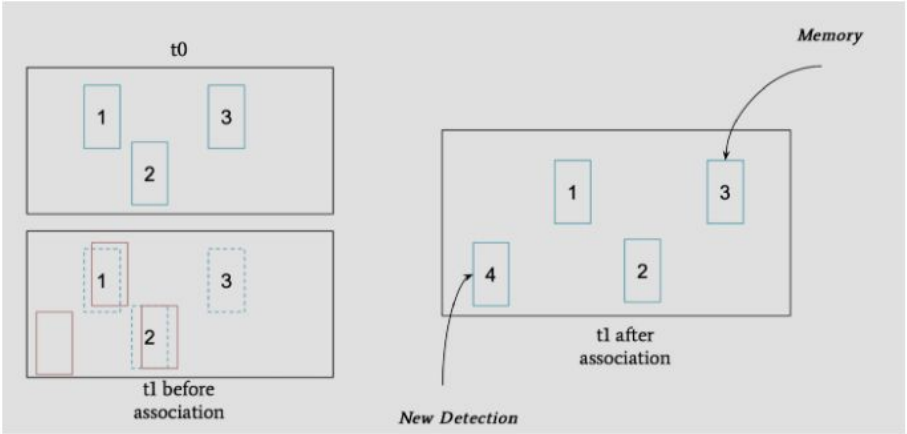


COCO(Common Object in Context) Dataset:

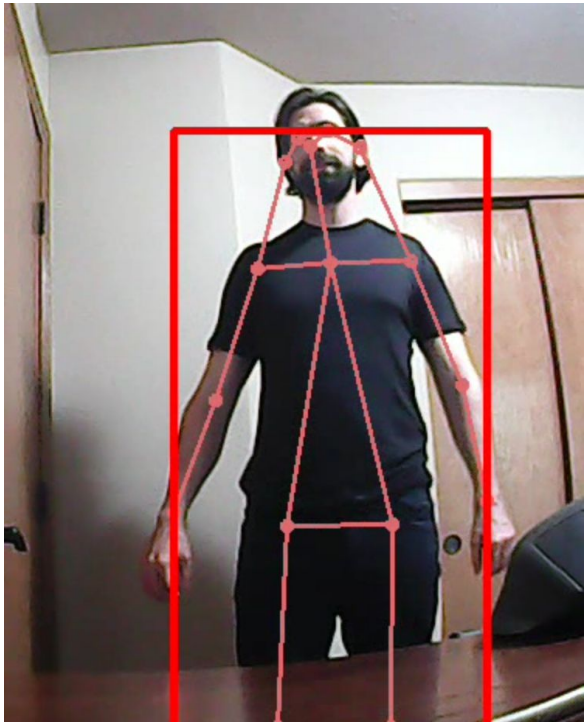
- 121K+ images with 262K+ humans with 1.8M+ body parts

Body Frame - The Hungarian Algorithm





Body Part detection - Output



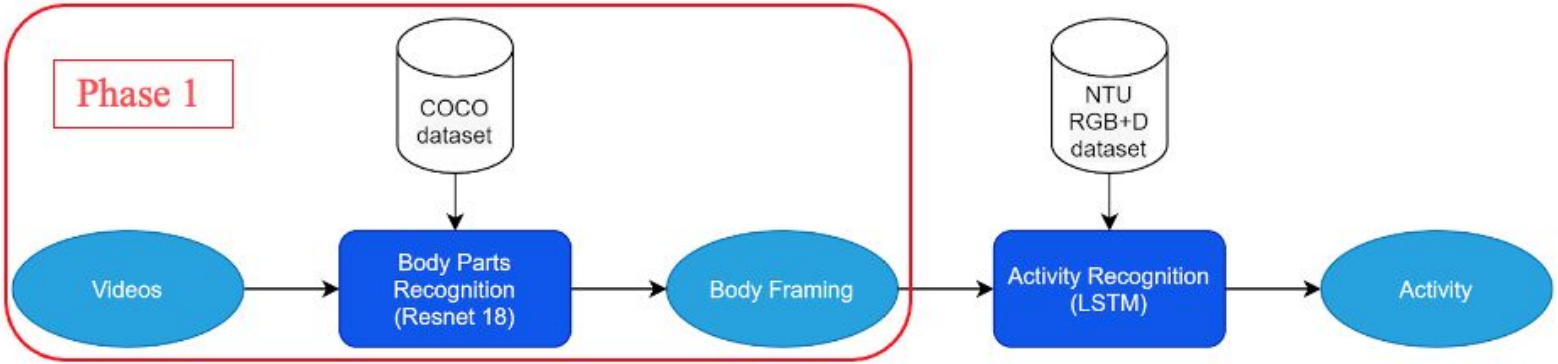
Activity Recognition - Model & Data

Input:

- Body Pose data with object detection from Phase 1
- NTU Dataset

Output

- Action Label (drink_eat; sit_squat; phone; walk; selfie)



Activity Recognition - NTU Dataset

Action	Attention iteration #1	Attention iteration #2
(1) Taking a selfie		
(2) Pointing to something		
(3) Kicking other person		

Fine Tuning Models

For Body Pose Detection (ResNet18)

- Image shape: 224 x 224 ; 368 x 368; 256 x 256
- Batch size: 64
- IoU: 0.5 - 0.95
- Area: all medium, large
- maxDets: 20
- Optimizer: Adam

For Action Recognition (LSTM)

- Batch size: 32 - 96
- Windows: 9
- Dropout: 0.2 - 0.4
- Optimizer: RMS prop
- Layers: 128 x 64

Validation AP: 0.45

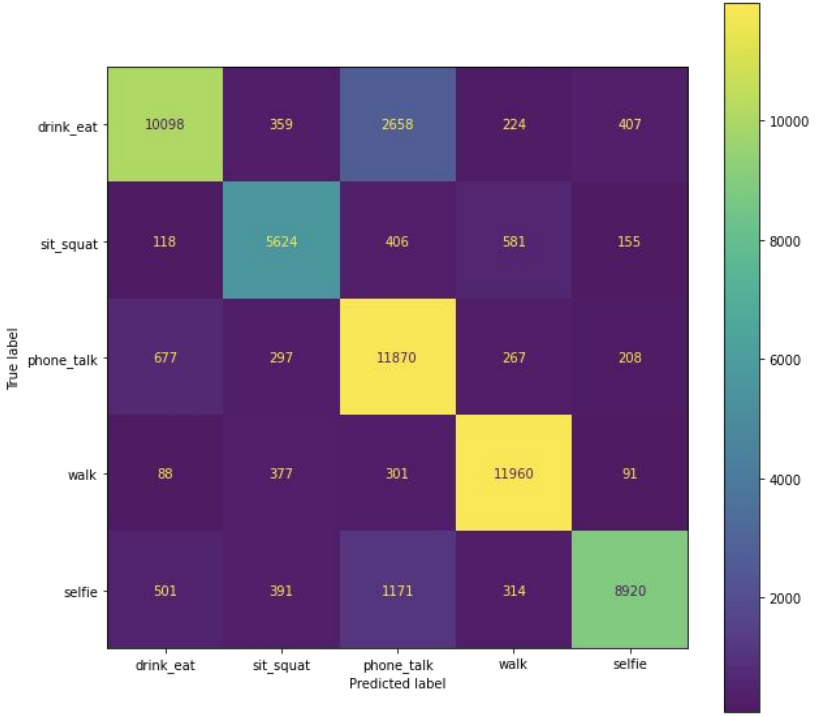
Validation AR: 0.60

Validation Loss: 0.52

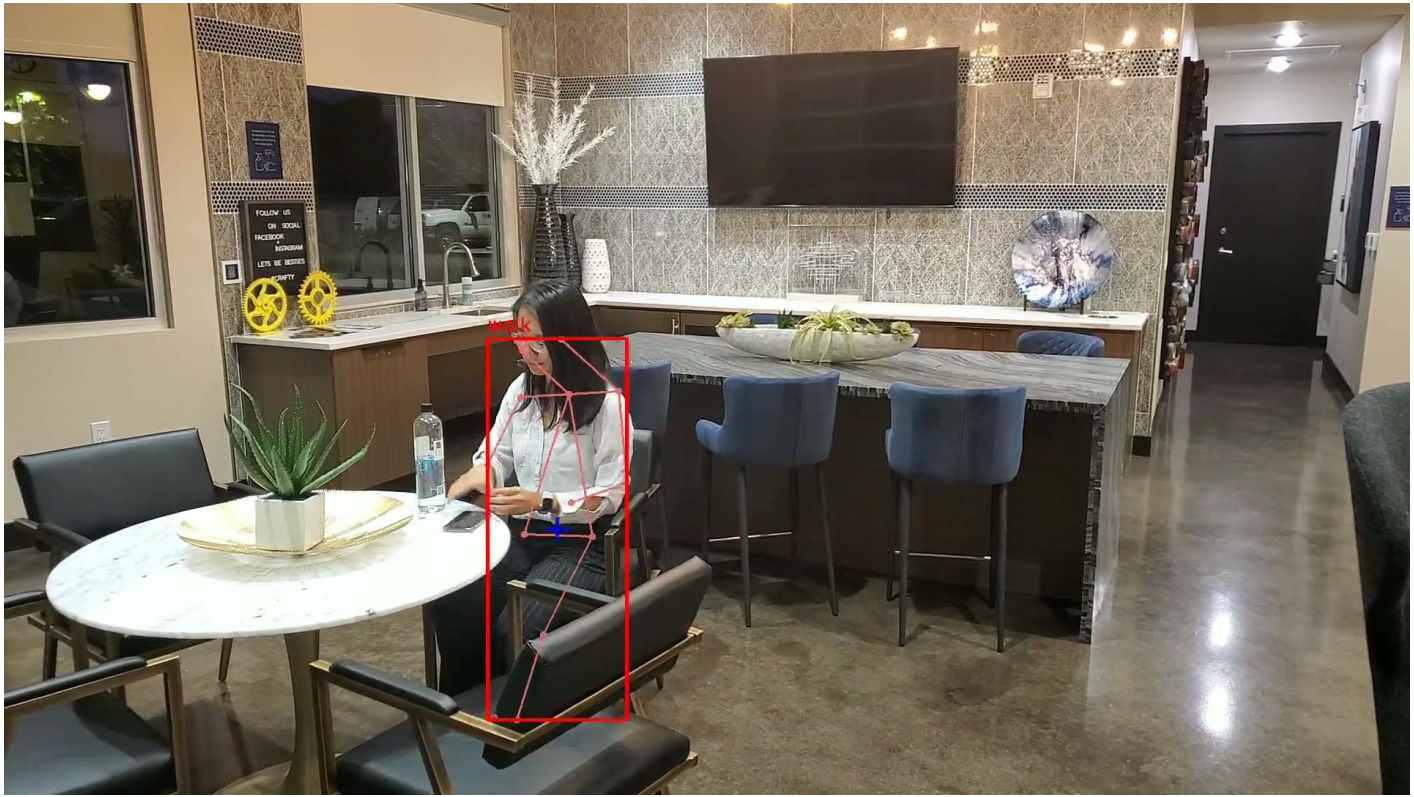
Validation Accuracy: 0.84

Activity Recognition - Output

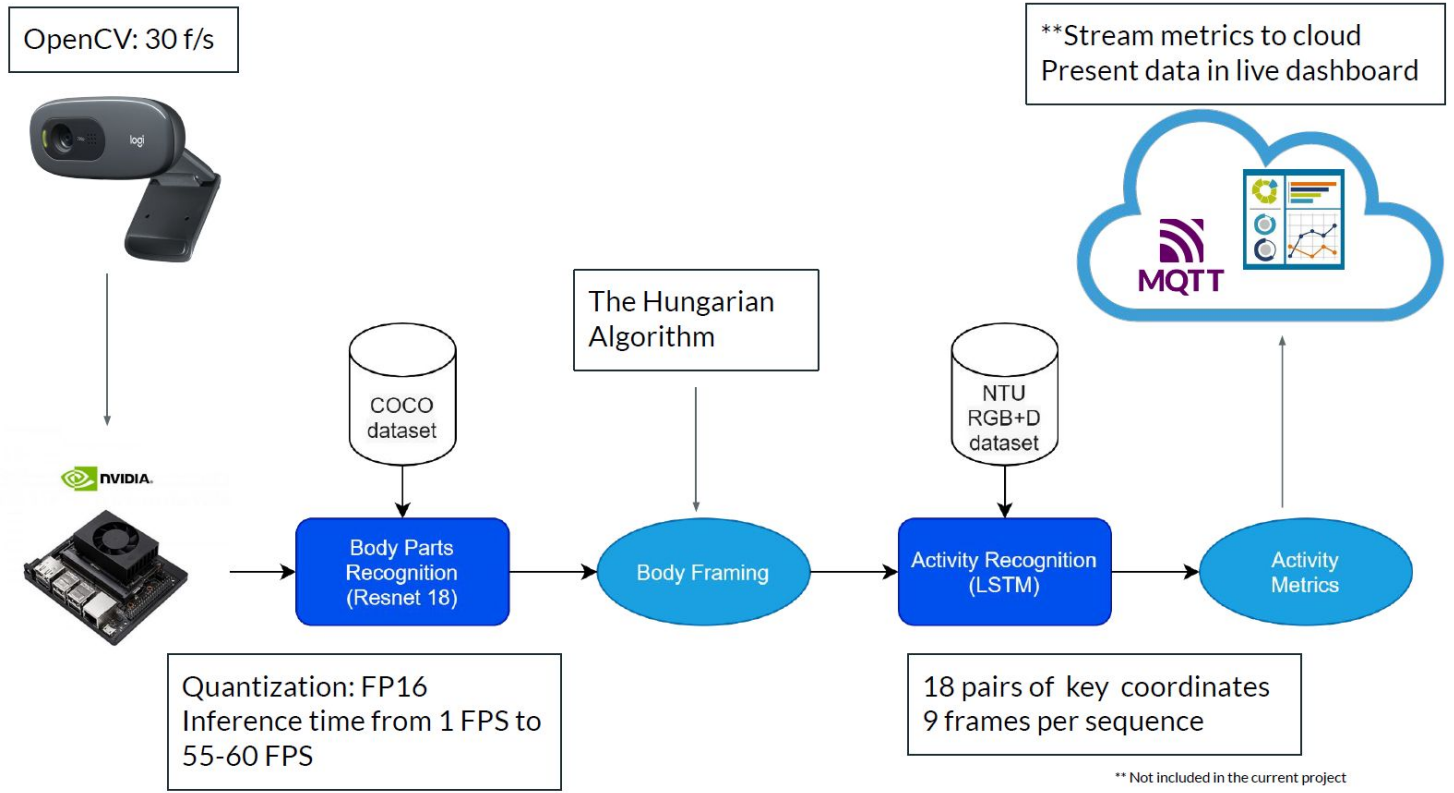
	precision	recall	f1-score	support
drink_eat	0.88	0.73	0.80	13746
phone_talk	0.72	0.89	0.80	13319
selfie	0.91	0.79	0.85	11297
sit_squat	0.80	0.82	0.81	6884
walk	0.90	0.93	0.91	12817
accuracy			0.83	58063
macro avg	0.84	0.83	0.83	58063
weighted avg	0.84	0.83	0.83	58063



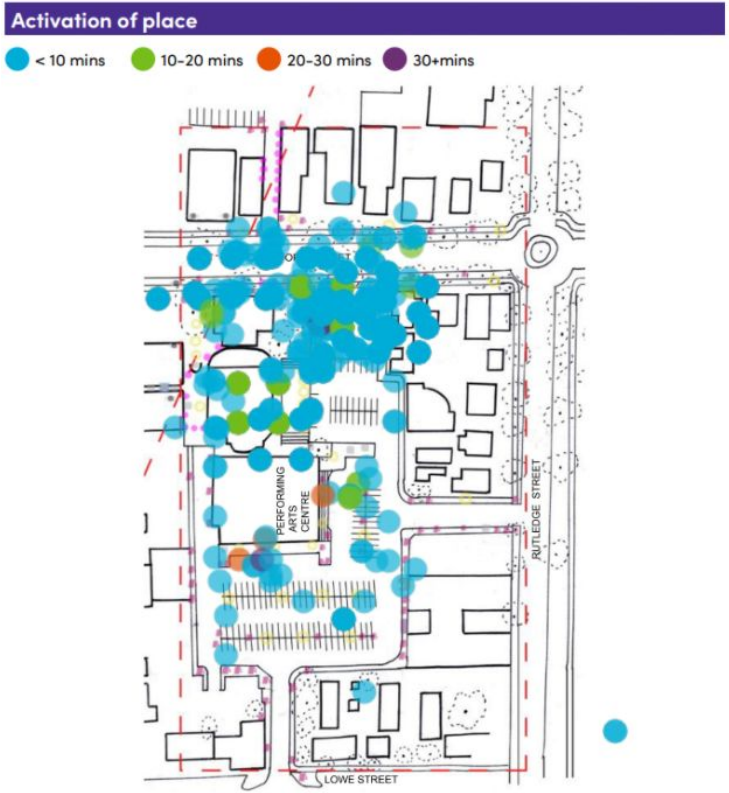
Simulation (Demo)



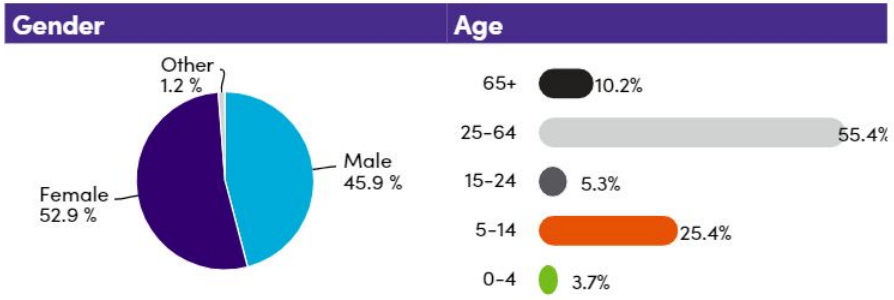
Solution Architecture



Scope for future work



- Demographic (Age/Gender) tracking
- Emotion capture
- Time Span at different places of park
- Streaming data to a dashboard
- Implement controls for privacy



Thank you!