Food Intake Activity Recognition Based on Privacy-Preserving mmWave Radars

Yi-Hung Wu (antasid@gapp.nthu.edu.tw)

Advisor: Cheng-Hsin Hsu

Networking and Multimedia Systems Lab,

Department of Computer Science, National Tsing Hua University

Outline

- Introduction
- Goal & Challenges
- Related Work
- Proposed Solutions
- Dataset
- Global Model Evaluations
- Leave-one-out Model Evaluations
- Conclusion & Future Work

A Smart Home with Heterogeneous Sensors

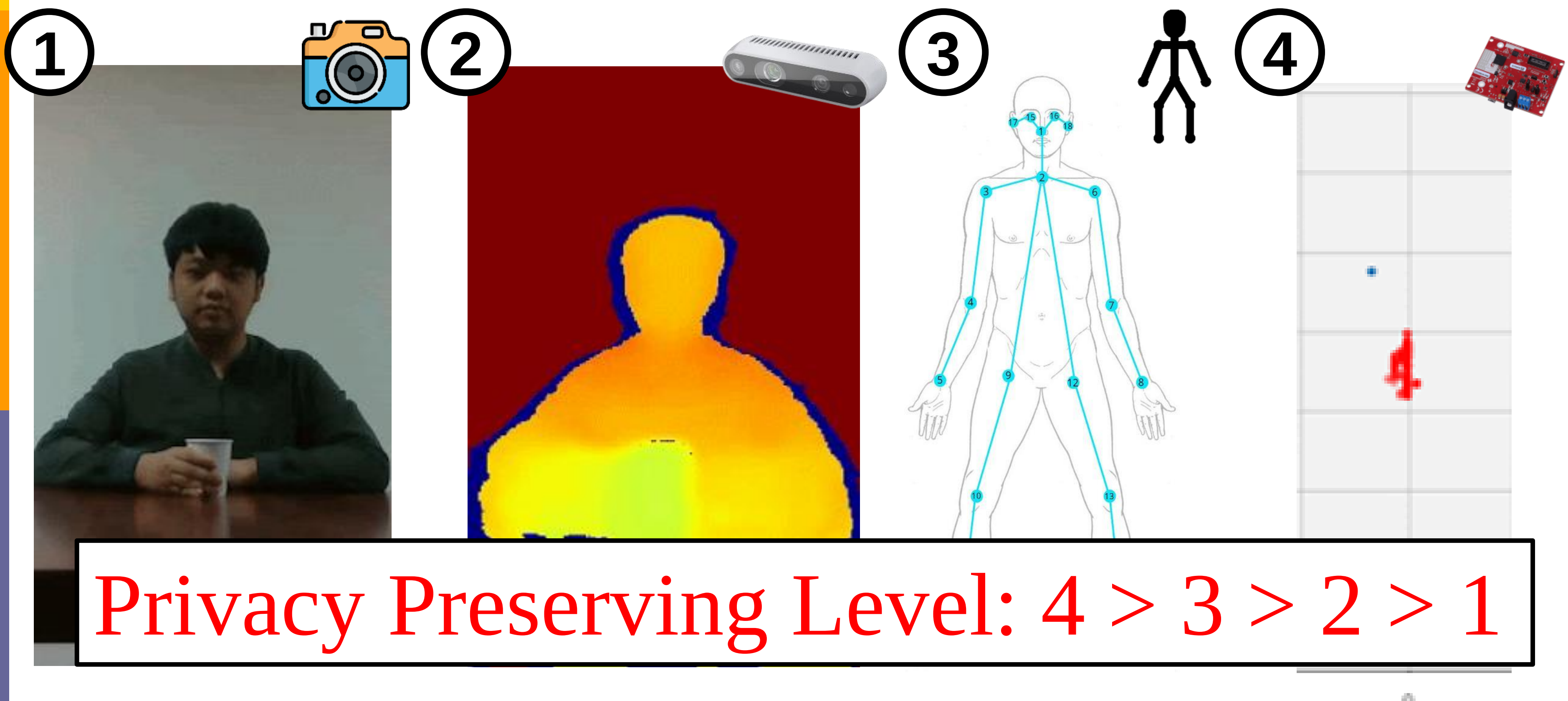
Applications for Food Intake Activity Recognition

- Diet Control
  - Automatic monitoring
  - Fasting management

- Telecare
  - Meal recording & reminder
  - Medication monitoring

- Smarthome
  - Eating behavior prediction
  - Food management

The Privacy Issue of Sensor Data



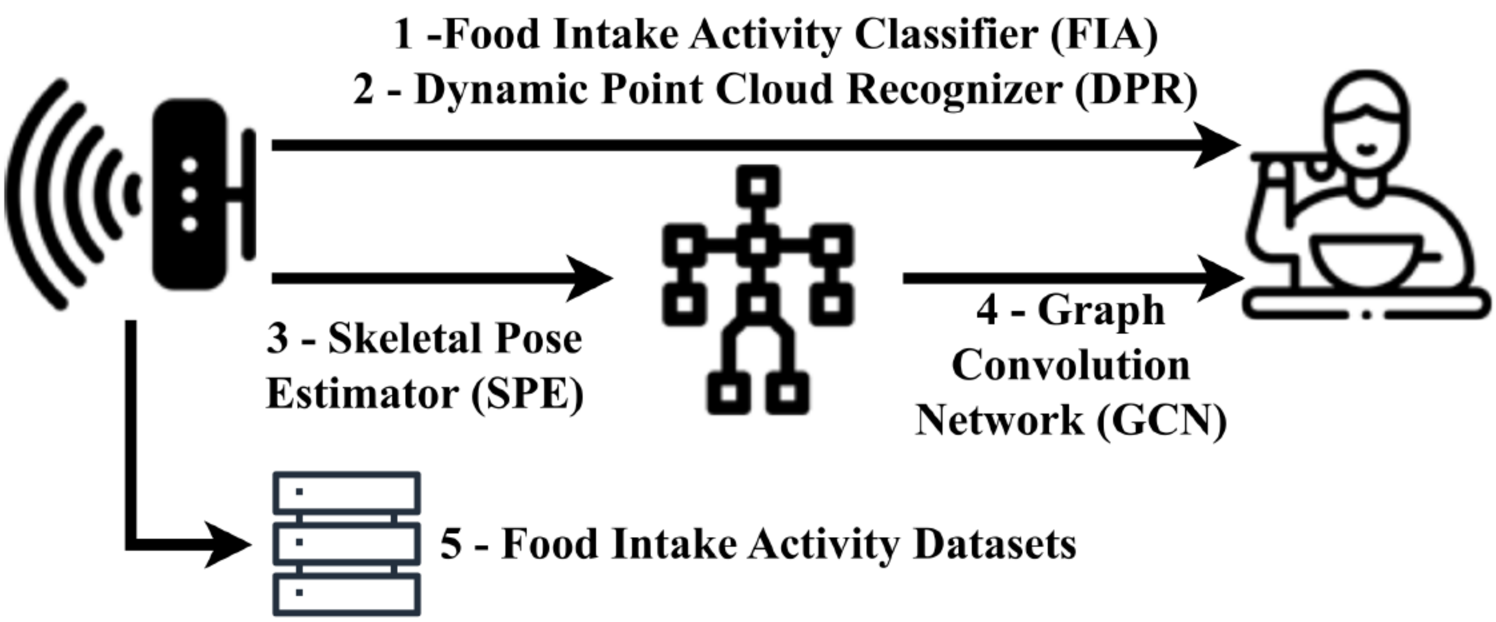
Privacy Preserving Level: 4 > 3 > 2 > 1

Outline

- Introduction
- Goal & Challenges
- Related Work
- Proposed Solutions
- Dataset
- Global Model Evaluations
- Leave-one-out Model Evaluations
- Conclusion & Future Work

Goals

- We want to detect:
  - "When" the person is eating/drinking
  - "How" the person is eating/drinking



Challenges

- Sensors
  - The sparsity of the mmWave point clouds
  - The sensitivity of the mmWave radar
- Datasets
  - No public dataset that focuses on human food intake activity
  - No mmWave radar dataset with multiple sensors data

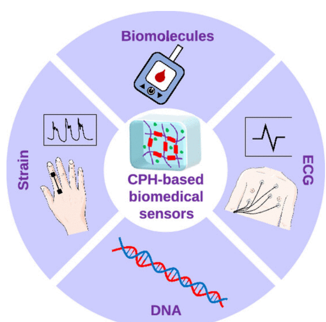
Outline

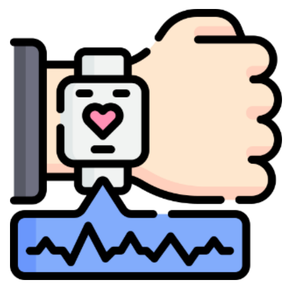
- Introduction
- Goal & Challenges
- Related Work
- Proposed Solutions
- Dataset
- Global Model Evaluations
- Leave-one-out Model Evaluations
- Conclusion & Future Work

Fine-Grained Activity Recognition

Wearable Sensors

- Bioelectric sensors
  - Electromyography (EMG) sensors [1]
  - Electroencephalography (EEG) sensors [2]

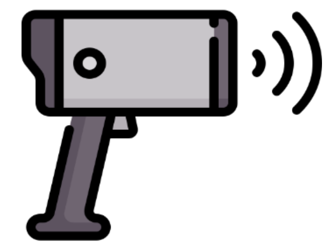




- Inertial sensors
  - Smartwatches [3]
  - Smartphones [3, 4]

In-situ Sensors

- Vision-based sensors
  - RGB camera
  - Depth camera
  - IR camera



- Radio Frequency (RF) sensors
  - WiFi [5]
  - mmWave radar [6]

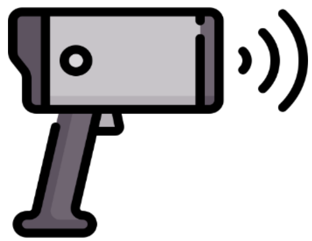
They require subjects to remember carrying with them

We use 3D mmWave radar to achieve higher accuracy while preserving privacy

[1] A.Moin et al. 2021. A wearable biosensing system with in-sensor adaptive machine learni
[2] A. Sa ... ectroencephalography
[3] G. W ... sing activities of daily
[4] N. A ... n smartphone sensor d
[5] L. Gu ... ognition
[6] S. Bhalla et al. 2021. Imu2doppler: Cross-modal domain adaptation for doppler-based acti

Skeletal Pose Estimation

- RGB-based
  - G-RMI [1]
  - DeepCut: Multi Person Pose Estimation [2]

- mmWave-based approaches
  - mmPose-NLP [3]
  - MARS [4]

Food Intake Activity Datasets

- Activities with Rich-Media Sensors
  - RGB-based dataset contains plenty of activities
    - NTU-RGBD [5]
    - Kinetics [6]
  - IMU/RF sensors contain only coarse-grained activities
- Food-Intake Activities with wearable Sensors

There is no public dataset for Food-Intake Activities with mmWave radars

[1] G. Papandreou et al. 2017. Towards accurate multi-person pose estimation in the wild

[2] L. Pishchulin et al. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation

[3] A. Sengupta and S. Cao. 2022. mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars.

[4] S. An and U. Ogras. 2021. Mars: mmwave-based assistive rehabilitation system for smart healthcare.

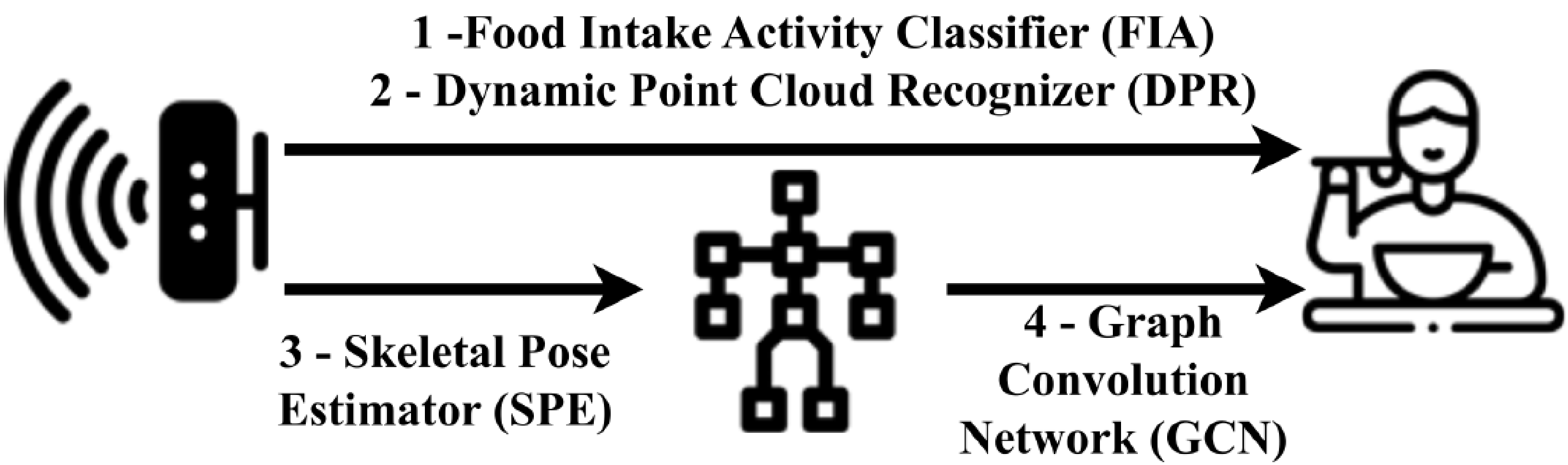
[5] A. Shahroudy et al. 2016. Ntu rgb+ d: A large scale dataset for 3D human activity analysis.

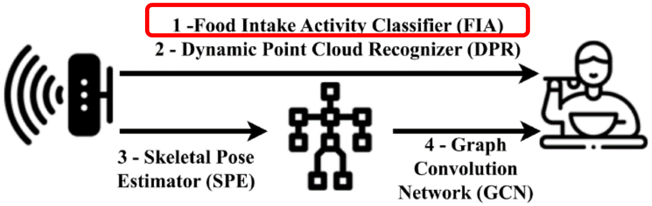
[6] W. Kay et al. 2017. The kinetics human action video dataset.

Outline

- Introduction
- Goal & Challenges
- Related Work
- Proposed Solutions
- Dataset
- Global Model Evaluations
- Leave-one-out Model Evaluations
- Conclusion & Future Work

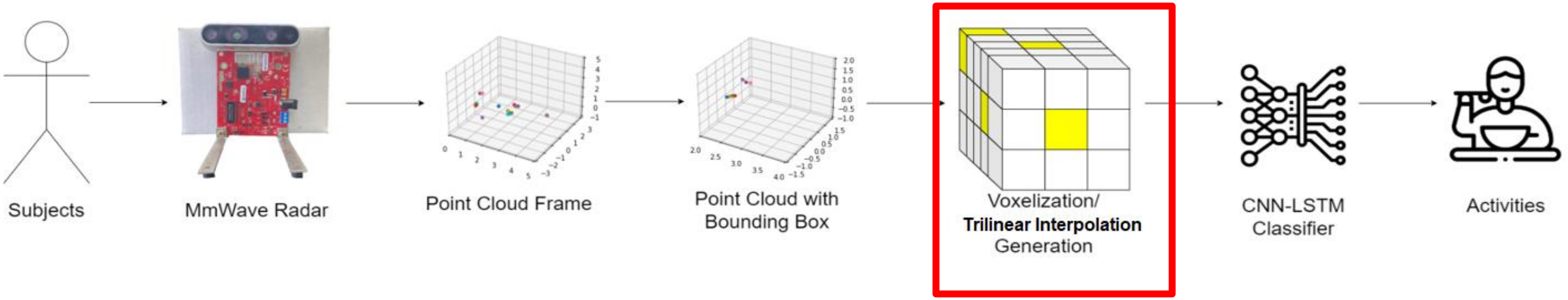
Overview

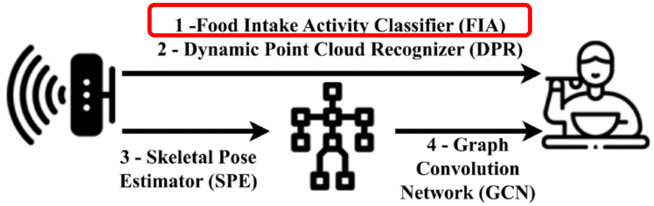

1 -Food Intake Activity Classifier (FIA)
2 - Dynamic Point Cloud Recognizer (DPR)
3 - Skeletal Pose Estimator (SPE)
4 - Graph Convolution Network (GCN)



FIA: End-to-End Voxelization Pipeline

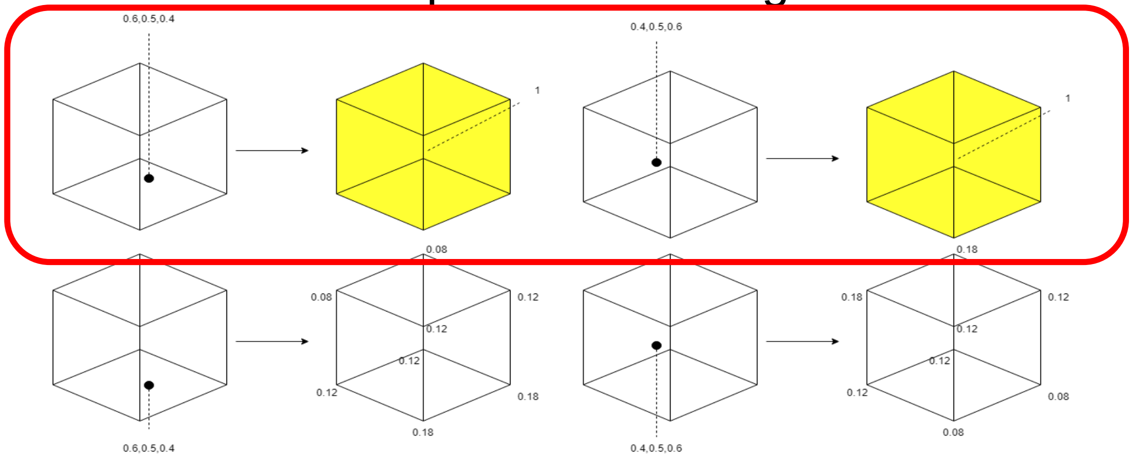
- The **voxelization** method is used to make mmWave radar point cloud trainable
- We proposed bounding box and trilinear interpolation method to improve the performance of the classifier for fine-grained actions
- A neural network classifier is proposed to recognize the activity

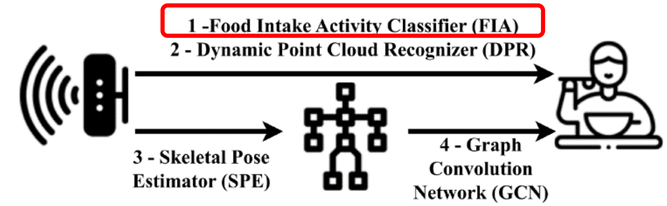




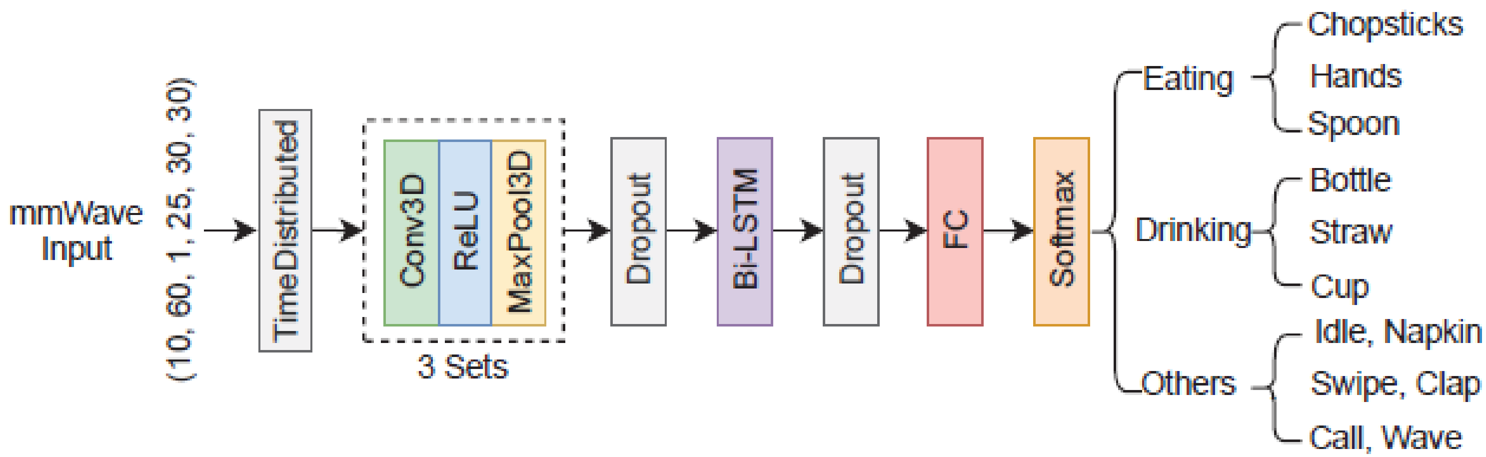
FIA: Trilinear Interpolations

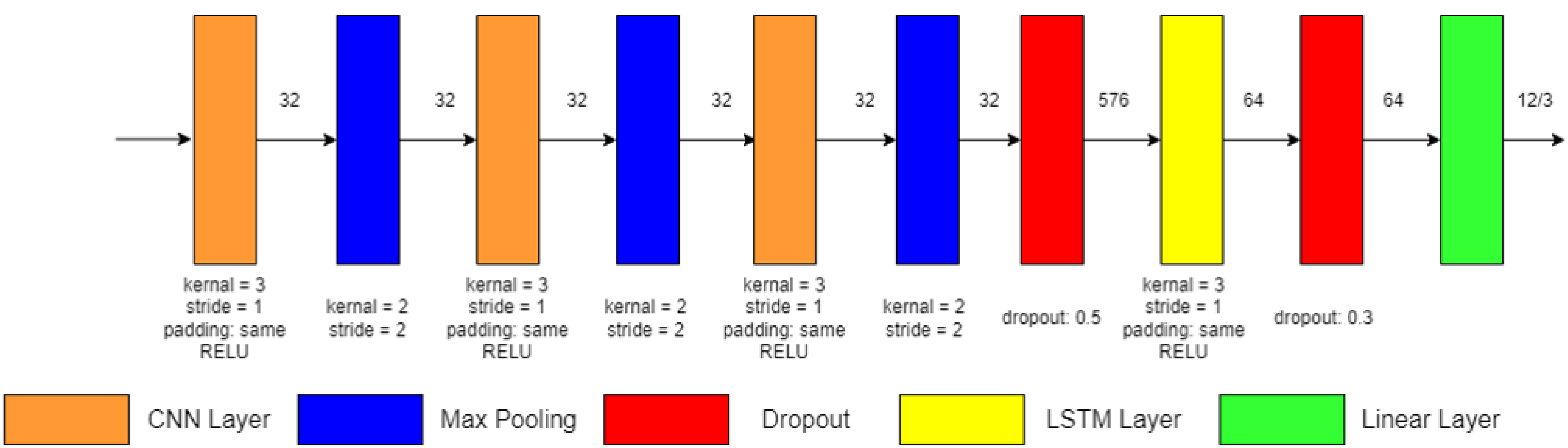
- Original voxelization process:
  - An element representing a cube in the bounding box
  - If a point in a cube, the element that represents the cube +1
- Our new method:
  - An element representing a vertex of a cube
  - If there is a point in a cube, 8 vertices of the cube will get a weight value, and the sum of the weight is 1
- Difference:
  - We can know the difference when the points is moving in the cube
  - More points is nonzero, which having more information for the classifier

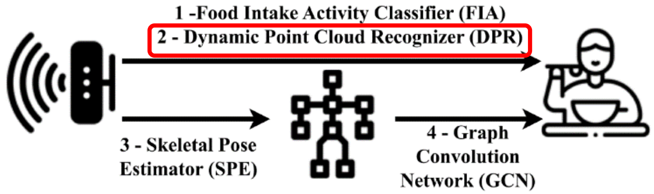




FIA: Neural Network Structure

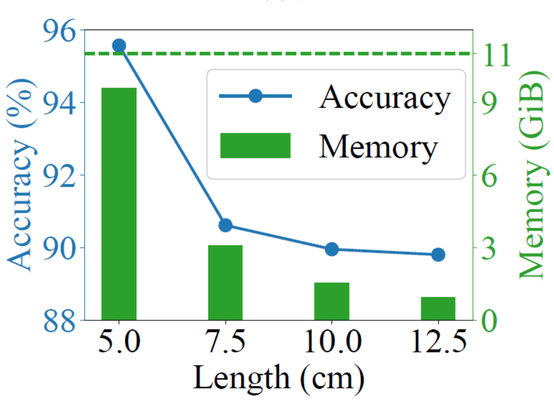


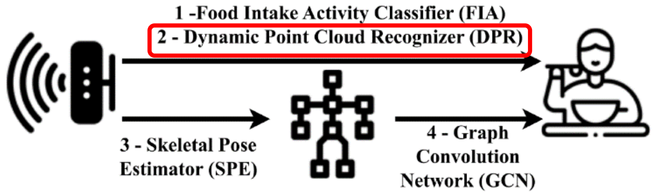




DPR: Introduction

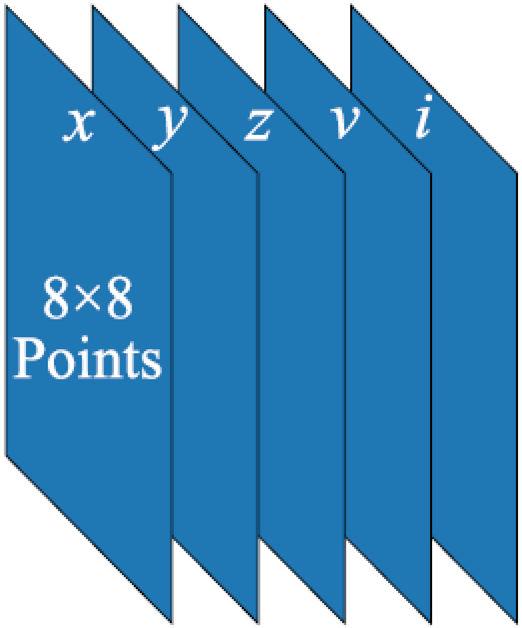
- Voxelization has two main disadvantage
  - The resolution affects the accuracy
  - Huge memory consumption
- DPR's main idea
  - Directly using point cloud data
  - Taking velocity, and intensity into consideration
- End-to-end activity classifier

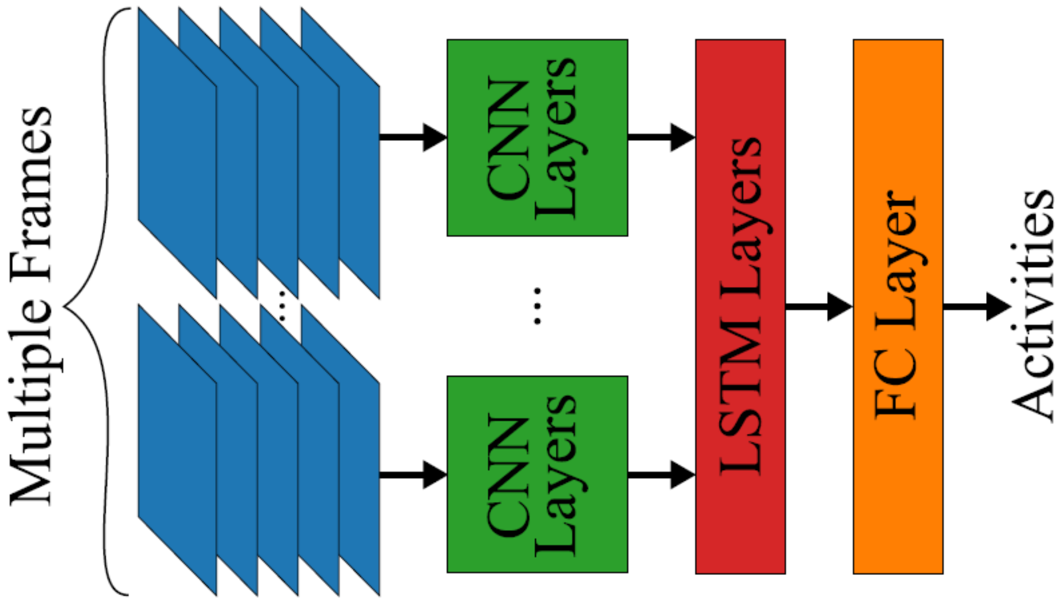


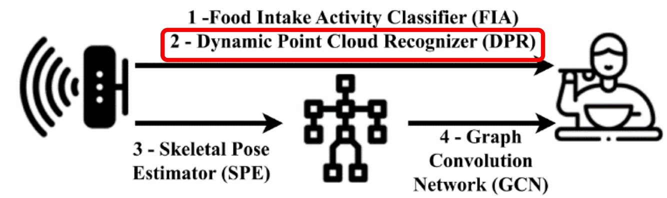


DPR: Model Structures

- Maximum 64 points per frame → Zero-padding
- 5 channels of input: X, Y, Z, velocity, intensity
- Multiple frames as temporal features→ LSTM layers

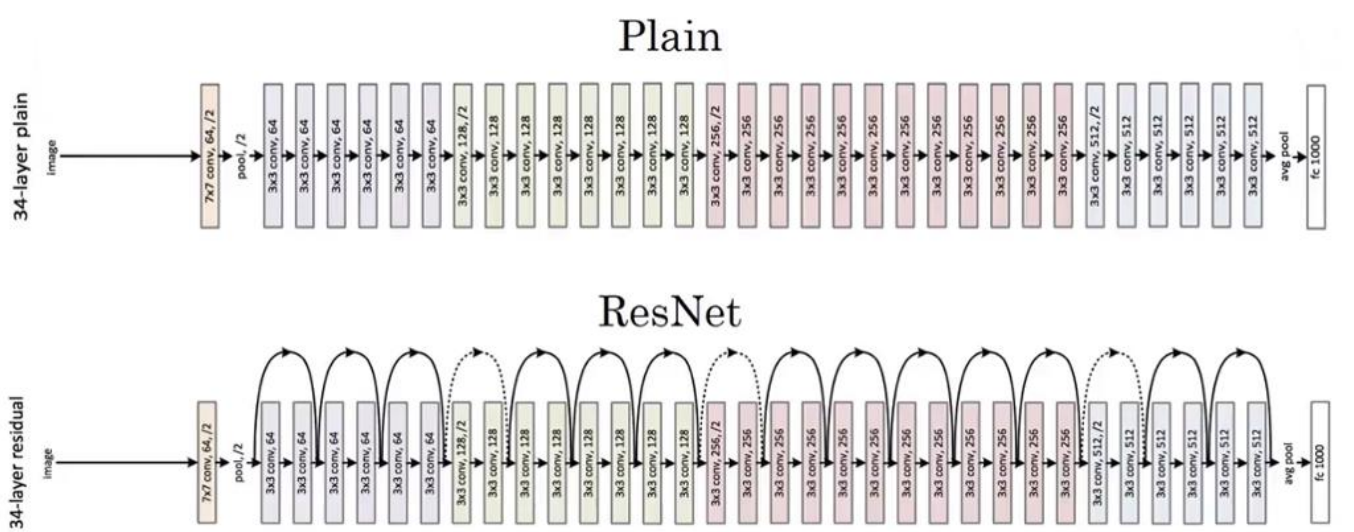






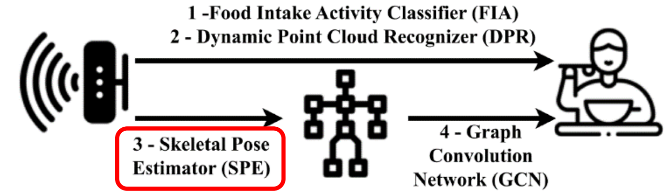
DPR: Adjusting CNN Layers

- Our proposed DPR uses more advanced and widely adopted network models
  - AlexNet
  - GoogLeNet
  - ResNet

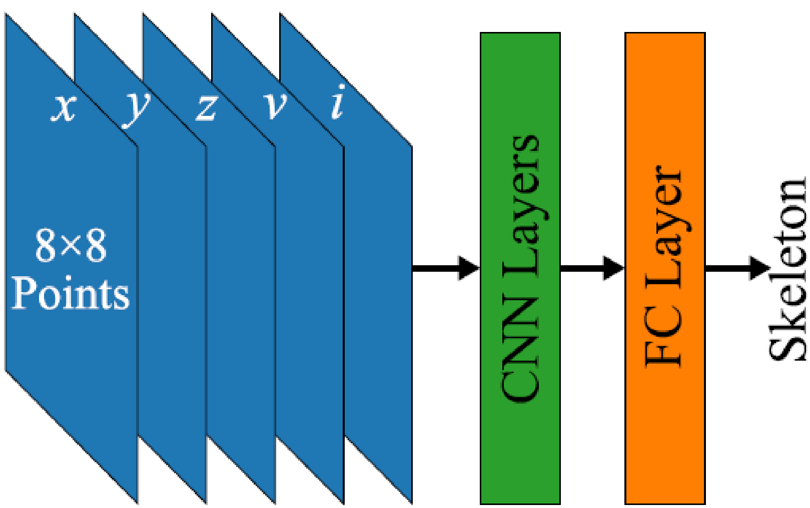


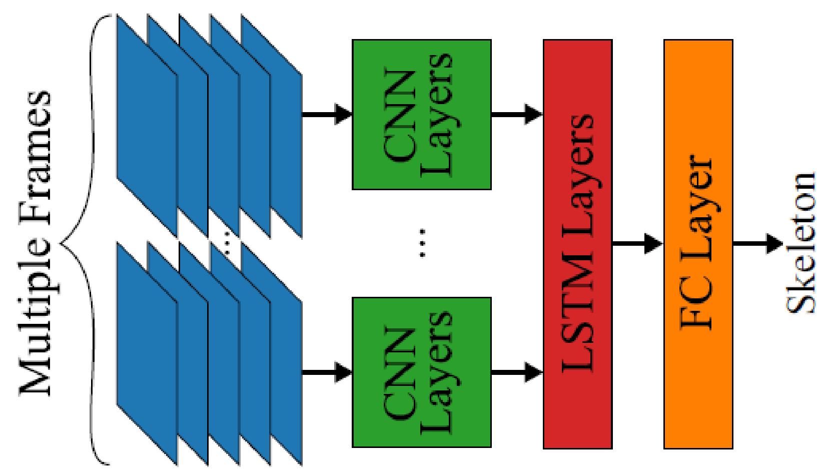
[1] K. He et al. 2015. Deep Residual Learning for Image Recognition

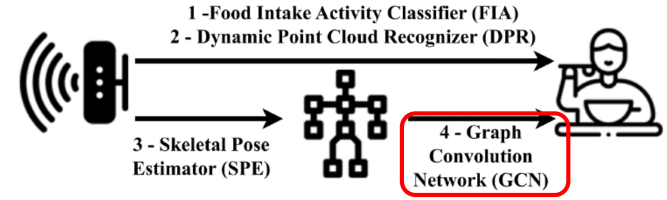
SPE: Introduction



- Same structure with DPR
- SPE: single frame skeleton estimator
- SPE+: skeleton estimator based on multiple frames’ data

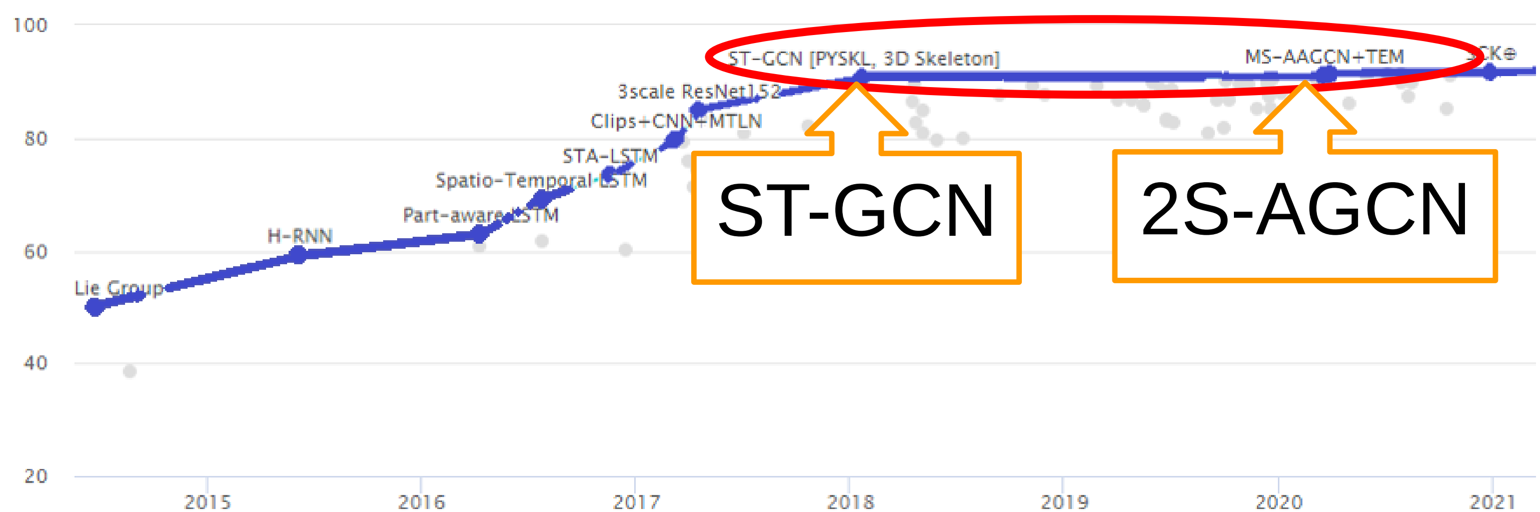




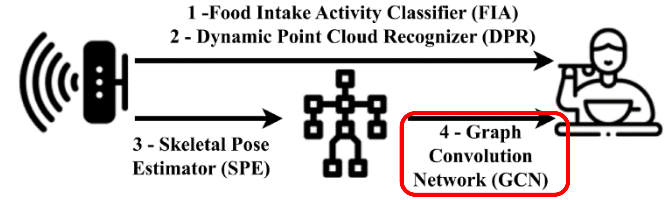


GCN: Introduction

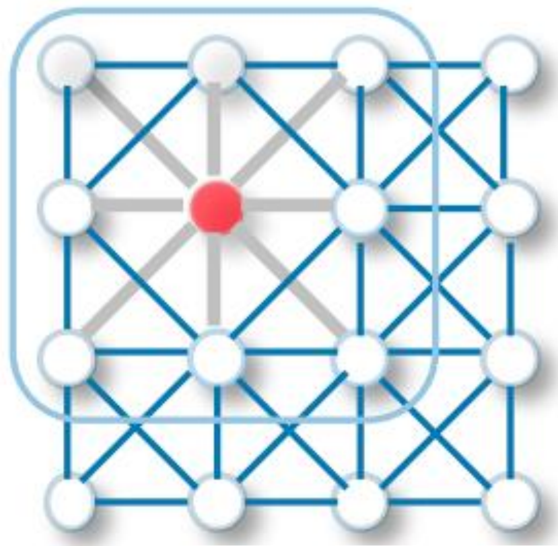
- Graph convolution network classifier is one of the best solution for skeleton data
- GCN classifier Implemented
  - ST-GCN
  - 2S-AGCN

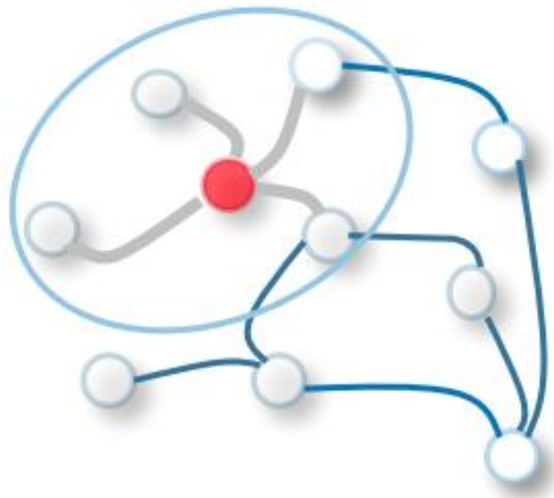


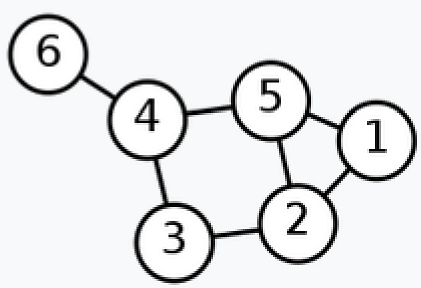
[1] Source: Paperwithcode; https://paperswithcode.com/task/skeleton-based-action-recognition



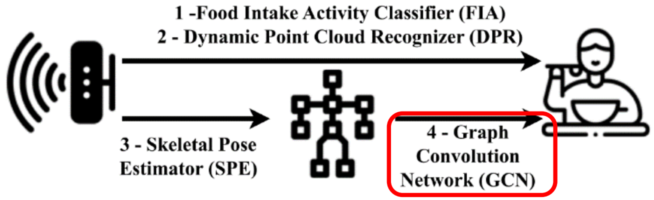
GCN: Graph Convolution Networks





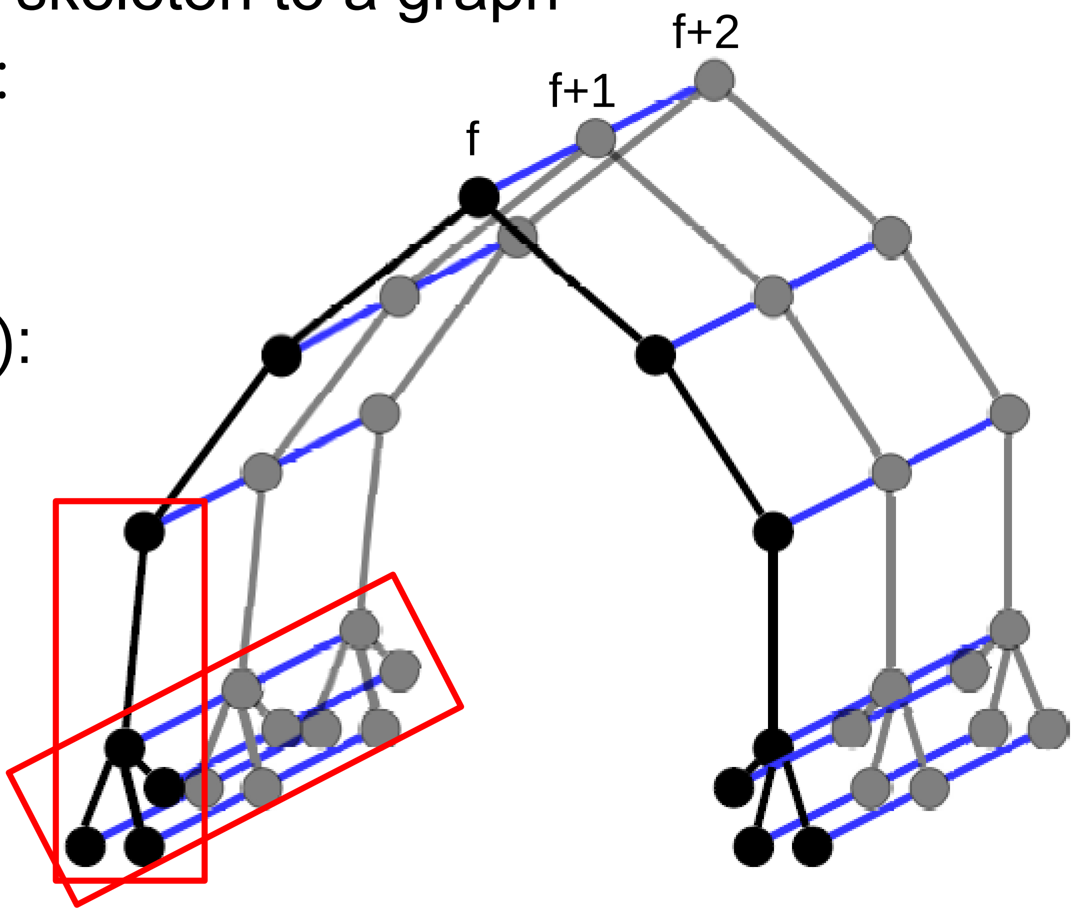
Labeled graph	Adjacency matrix
	$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$

[1] Understanding Graph Convolutional Networks for Node Classification; https://towardsdatascience.com/understanding-graph-convolutional-networks-for-node-classification-a2bfdb7aba7b



GCN: Graph Construction

- Merge several frame's skeleton to a graph
- Spatial edges (Black) : Spacial joints at the same time
- Temporal edges (Blue): Same joints at the different time

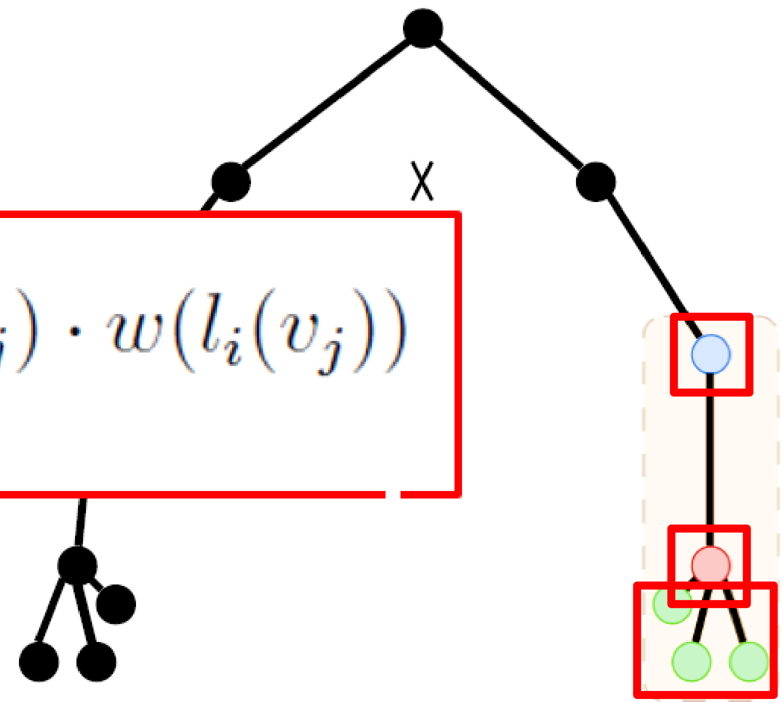


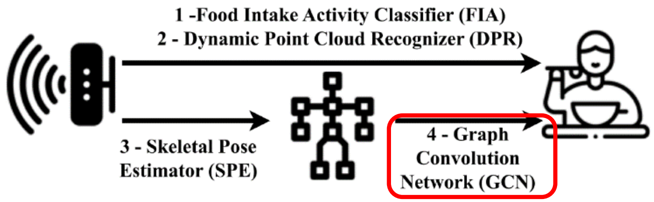
GCN: Graph Convolution

- The convolution area is fixed: all of the vertexes with a distance of 1
- P
  - center

$$f_{out}(v_i) = \sum_{v_j \epsilon B_i} \frac{1}{Z_{ij}} f_{in}(v_j) \cdot w(l_i(v_j))$$

  - Subset 2 (blue): inner subset
  - Subset 3 (green): outer subset





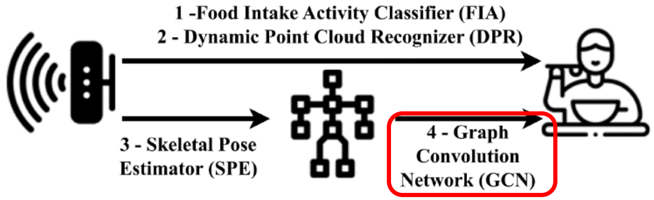
GCN: Graph Feature Extraction

- ST-GCN:
  - A matrix is the linking matrix including self-edge
  - M matrix is a trainable matrix

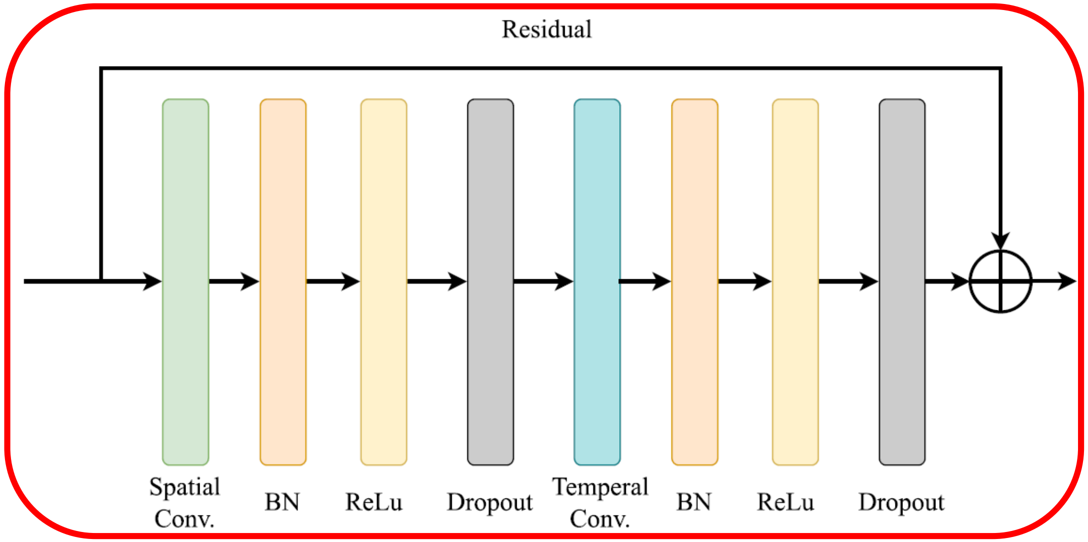
$$f_{out} = \sum_{k}^{K_v} W_k (f_{in} A_k) \odot M_k$$

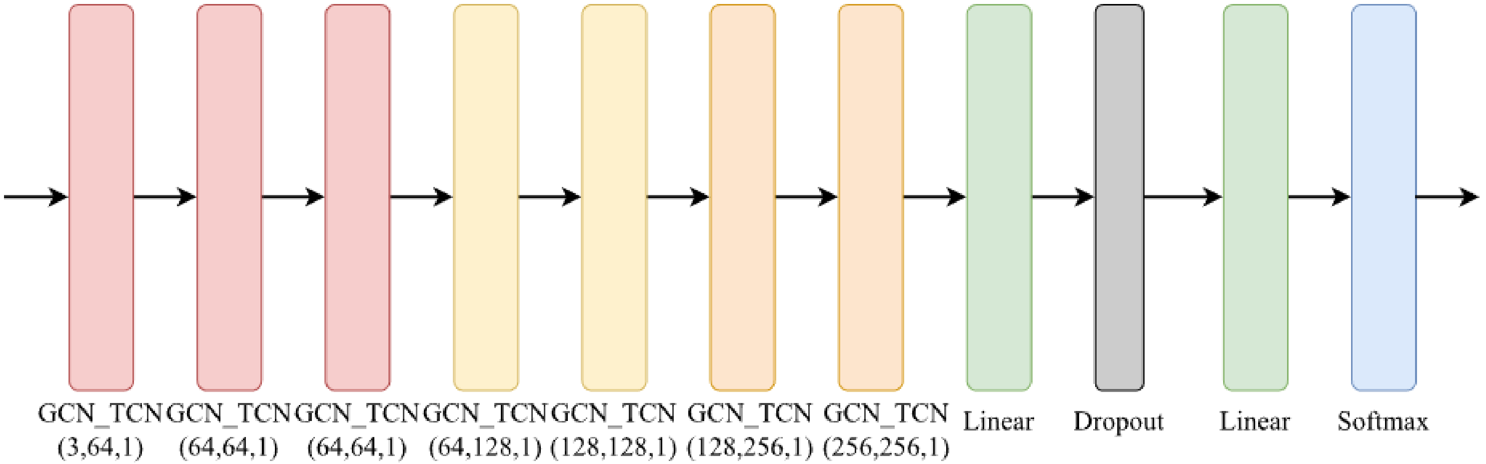
- 2S-AGCN:
  - B matrix is similar to M matrix
  - C matrix is a normalized matrix, recording the similarity of vertexes

$$f_{out} = \sum_{k}^{K_v} W_k f_{in} (A_k + B_k + C_k)$$



GCN: Model Structures





Outline

- Introduction
- Goal & Challenges
- Related Work
- Proposed Solutions
- Dataset
- Global Model Evaluations
- Leave-one-out Model Evaluations
- Conclusion & Future Work

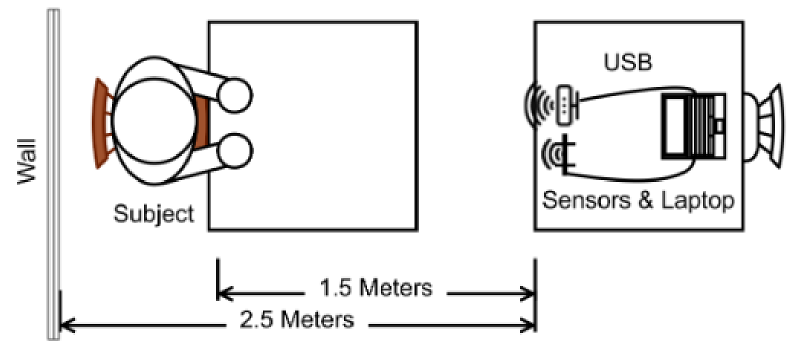
Motivation

- Human activity recognition problems
  - Coarse-grained activities [1, 2]
  - Fine-grained activities[3]
  - Food intake activities[4]
- There is no fine-grained food intake activity recognition dataset with privacy-preserving sensors such as mmWave radar
- We generate the very first Food Intake Activity dataset with different privacy sensitivity sensors.

[1] Y. Huang et al. 2022. Activity Recognition Based on Millimeter-Wave Radar by Fusing Point Cloud and Range–Doppler Information.
[2] A. Logacjov et al. 2021. HARTH: A Human Activity Recognition Dataset for Machine Learning
[3] D. Anguita et al. 2013. A Public Domain Dataset for Human Activity Recognition Using Smartphones
[4] Y. Wu et al. 2022. AI-Assisted Food Intake Activity Recognition Using 3D mmWave Radars

Dataset Collection

- Environment Setup
  - The subject is 1.5 meters from the sensors
  - A wall is about 2.5 meters from the sensors
  - The table is 75 cm high
- Hardware Setup
  - Intel Realsense D435i RGB-D camera
  - TI IWR1443BOOST mmWave radar
- Software Setup
  - OS: Ubuntu 20.04
  - Pyrealsense 2 (librealsense) [1]
  - TI mmWave ROS Package [2]



[1] Intel® RealSense™ SDK 2.0. https://github.com/IntelRealSense/librealsense
[2] Leo Zhang. 2019. Github-radar-lab/ti_mmwave_rospkg. https://github.com/radar- lab/ti_mmwave_rospkg

Sensors

Sensor	mmWave Radar	RGB-D Camera
Laptop OS	Ubuntu 20.04	
Model	TI IWR1443BOOST	Intel RealSense D435i
Driver	TI mmWave rospkg	Pyrealsense2
SDK	TI mmWave SDK	Librealsense
Data Type	Dynamic point clouds	RGB/Depth video clips
Frame Rate	10 fps	30 fps

Description	Value	Description	Value
Starting frequency	77 GHz	Range resolution	4.4 cm
Bandwidth	3.44 GHz	Max range	3.95 m
Frame rate	0.1 s	Velocity resolution	7 cm
No. chirps per frame	32	Max velocity	1 m
No. TX antennas	4	Peak grouping	*False*
No. RX antennas	3	Clutter removal	*On/Off*

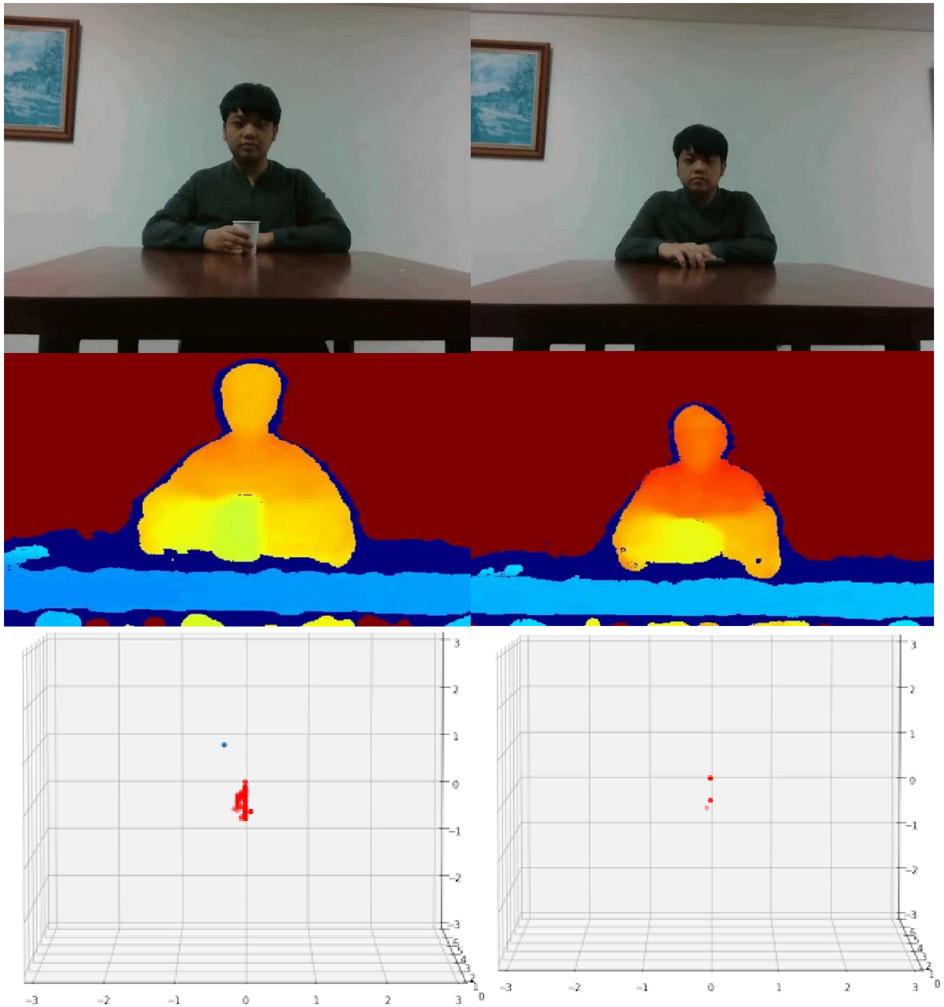
Dataset

- 12 classes of activities collected from 24 subjects
  - 6 food intake related
  - 6 other activities
- 2 different sensor's data, providing 4 types of data
- 2 different settings for mmWave radar (w/ and w/o clutter removal)
- A subject performs an activity for 30 times
- An activity sample is 4-seconds long
- In total, 19.2 hours of data is collected, with 5760 files, formed our dataset

Food Intake	Others
(a01) drinking tea with a cup	(a07) Idle
(a02) drinking tea with a bottle	(a08) picking up a call
(a03) drinking tea with a straw	(a09) cleaning one's mouth with tissue
(a04) eating burger with both hands	(a10) writing
(a05) eating fruits with a fork	(a11) reading
(a06) eating noodles with chopsticks	(a12) scrolling one's smartphone

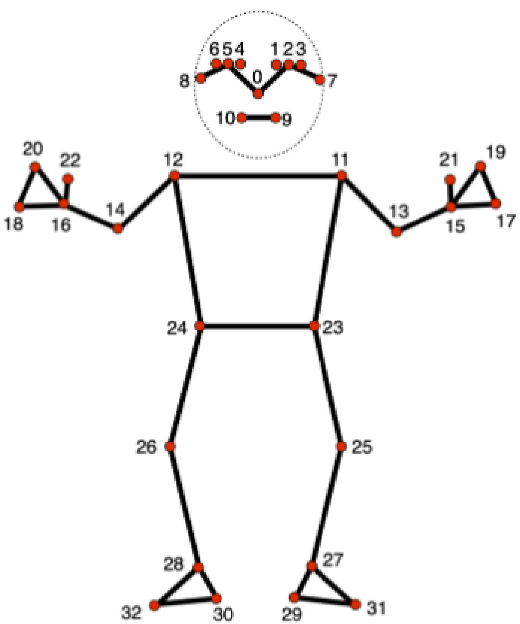
Dataset Sample Demo

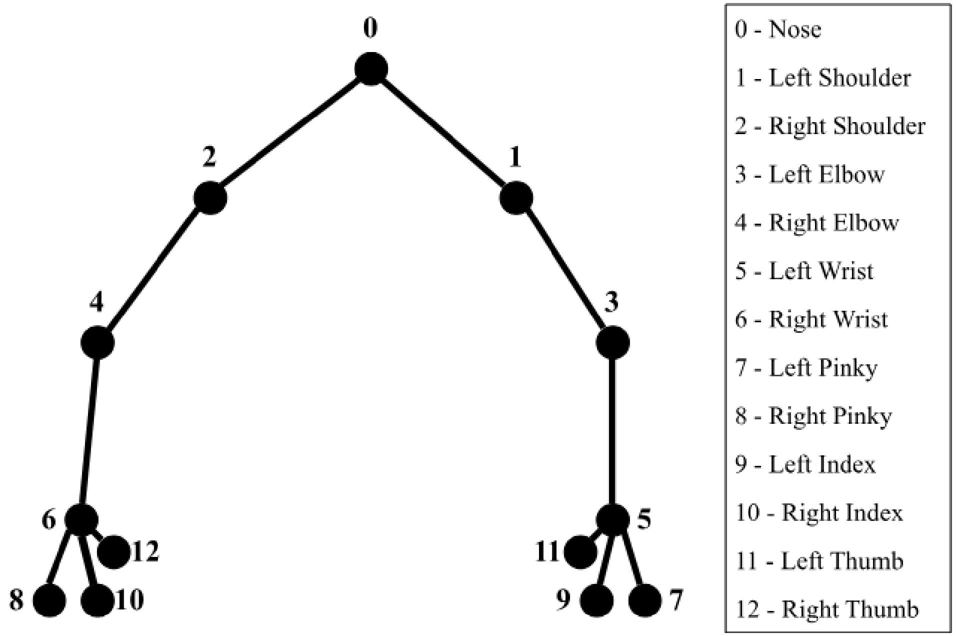
- 4 different data of the sample
  - RGB video
  - Depth map
  - Depth video
  - mmWave radar point cloud



Skeleton Generation

- Mediapipe Pose Model was utilized to synthesize human skeleton data
- Only 13 of 33 points are chosen to be our skeleton dataset



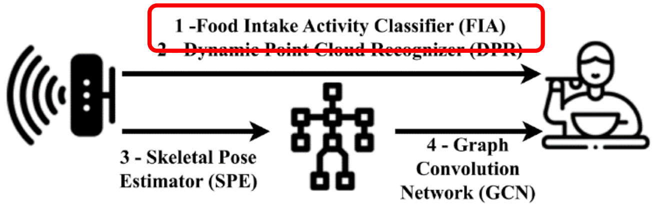


Outline

- Introduction
- Goal & Challenges
- Related Work
- Proposed Solutions
- Dataset
- Global Model Evaluations
- Leave-one-out Model Evaluations
- Conclusion & Future Work

Global Model Setup

- 80-20% random train-test split
- Data from each subject will appear in both the training and testing sets
- Simulate the scenario in the subjects have provided their data in advance



FIA – Experiment Setup

- Variants of FIA algorithms
  - FIA-D
  - FIA-B
  - FIA-V
- Preprocess parameters
  - temporal aggregation frames: $k \in \{1, 3, 5, 7, 9\}$
  - bounding box size: $(bX, bY, bZ) \in \{(2, 3, 3), (2, 3, 2), (2, 3, 1)\}$
  - resolution: $r \in \{10, 15, 20\}$
- Dataset
  - "When" dataset: 3 labels
  - "How" dataset: 12 labels
- Baseline: RadHAR

Food Intake	Others
drink with a cup	no activity
drink with a bottle	using smartphone
drink with straw	Phone call
eat with both hands (burger)	hand clap
eat with spoon	hand waving
eat with chopsticks	clean with tissue

FIA Detects Food Intake Activities With Good Performance

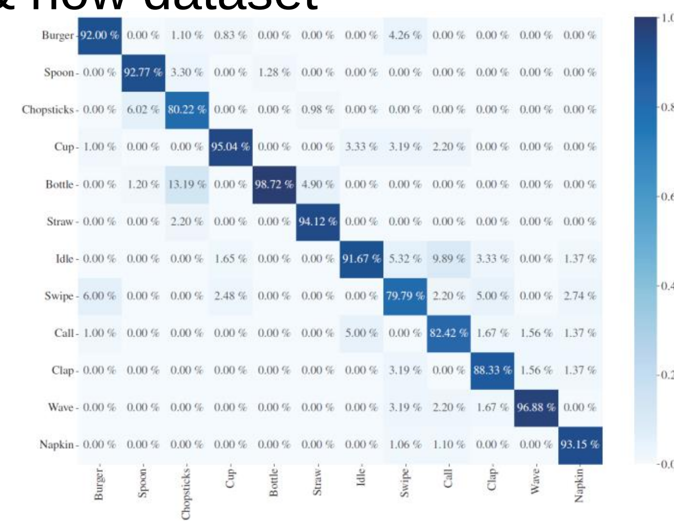
- FIA outperforms the SOTA voxelization solution in all varients

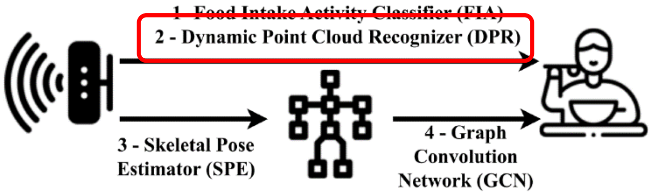
Algorithm	When Dataset	How Dataset
RadHAR	76.43%	12.17%
FIA-D	90.79%	68.81%
FIA-B	93.56%	72.77%
FIA-V	96.73%	91.49%

20.30% / 88.32%

- FIA shows high accuracy in both when & how dataset

Accuracy	Drinking	Eating	Others
Drinking	95.35%	2.13%	1.16%
Eating	2.71%	93.60%	0.46%
Others	1.94%	4.27%	98.38%



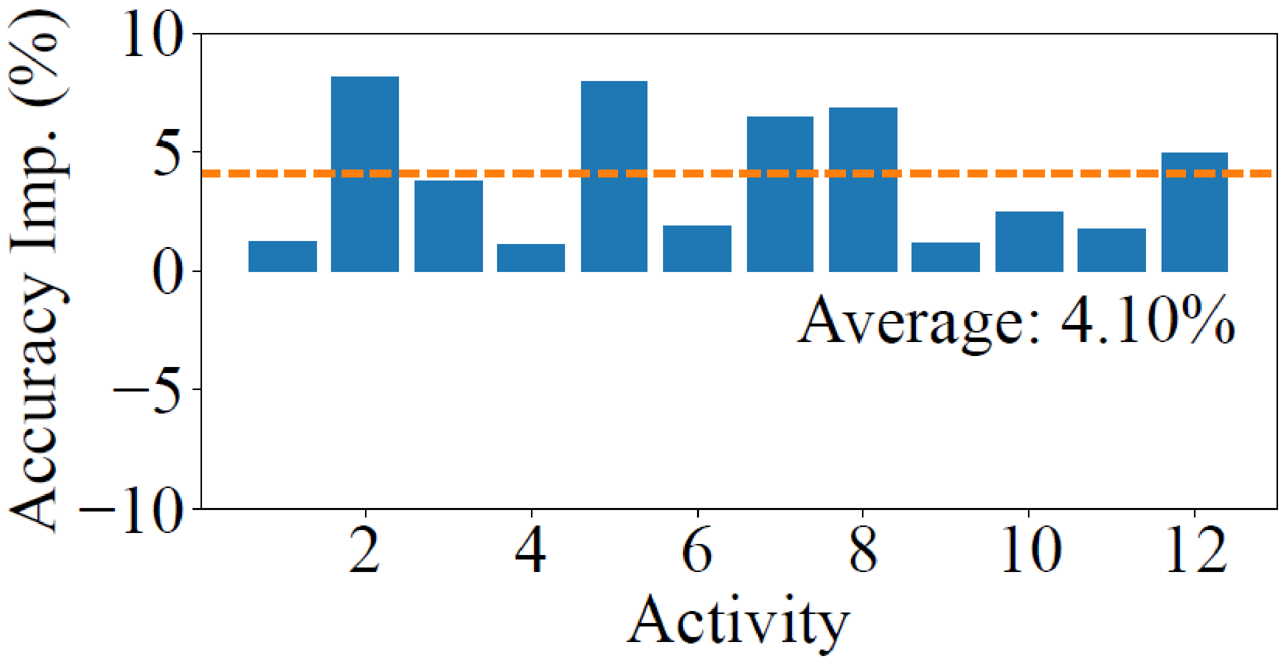


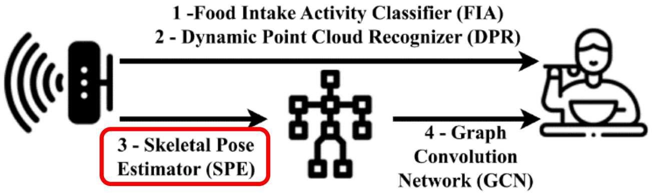
DPR - Experiment Setup

- Parameters
  - Output size (L): {**39**, 256, 576}
  - Number of LSTM layers (N) : {**1**, 2, 3}
  - Number of hidden LSTM states (H): {64, **128**, 256}
  - dropout rate (D): {0.1, **0.3**, 0.5}
  - Bidirectional LSTM (B) : {**true**, false}
  - Frames per sample (F): 40 (4 seconds)
- Dataset: “How” dataset (12 labels)
- Variants of DPR algorithm
  - AlexNet
  - GoogLeNet
  - ResNet
- Baseline: FIA

DPR Saves Memory But Also Improve The Accuracy

- FIA achieves a classification accuracy of 95.56%, while DPR achieves **99.66**%
- FIA utilizes 9817 MiB of GPU memory, while DPR only uses 2131 MiB, resulting in a 78.29% reduction
- Resnet algorithms achieved the best accuracy





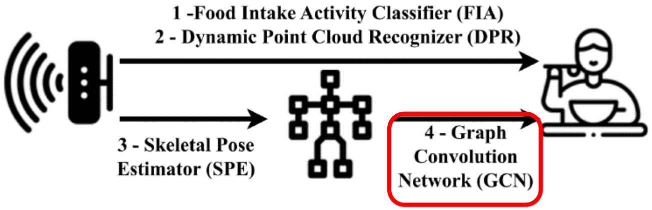
SPE – Experiment Setup

- Baseline
  - mmPose-NLP (NLP)
  - MARS
- Variants
  - AlexNet
  - GoogLeNet
  - Resnet-18, 34, 50
- Temporal aggregation frames of SPE+: {3, 5, 7, 9, 11}

SPE with ResNet-34 Is the Best Performance Variant

	NLP	MARS	AlexNet	GoogLeNet	ResNet-18	ResNet-34	ResNet-50
Nose	9.85 (±0.18)	8.28 (±0.17)	6.91 (±0.06)	4.73 (±0.05)	4.41 (±0.05)	**4.22 (±0.05)**	4.75 (±0.08)
L. Shldr	7.93 (±0.18)	6.66 (±0.10)	5.65 (±0.05)	3.91 (±0.04)	3.74 (±0.04)	**3.58 (±0.04)**	3.93 (±0.05)
R. Shldr	7.84 (±0.18)	6.58 (±0.11)	5.54 (±0.05)	3.86 (±0.04)	3.67 (±0.04)	**3.52 (±0.04)**	3.91 (±0.06)
L. Elbow	10.31 (±0.18)	8.68 (±0.15)	7.13 (±0.05)	4.84 (±0.04)	4.59 (±0.04)	**4.39 (±0.04)**	4.83 (±0.05)
R. Elbow	9.97 (±0.18)	8.38 (±0.16)	7.05 (±0.05)	4.97 (±0.04)	4.76 (±0.04)	**4.57 (±0.04)**	4.94 (±0.04)
L. Wrist	13.68 (±0.18)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6.55 (±0.06)
R. Wrist	14.02 (±0.18)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7.33 (±0.06)
L. Pinky	14.91 (±0.18)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7.13 (±0.06)
R. Pinky	15.79 (±0.18)	13.27 (±0.15)	11.47 (±0.07)	8.24 (±0.06)	8.08 (±0.06)	**7.74 (±0.06)**	8.28 (±0.06)
L. Index	14.91 (±0.18)	12.54 (±0.13)	10.50 (±0.07)	7.20 (±0.06)	7.00 (±0.05)	**6.67 (±0.05)**	7.21 (±0.06)
R. Index	15.68 (±0.18)	13.16 (±0.11)	11.41 (±0.07)	8.23 (±0.06)	8.06 (±0.06)	**7.73 (±0.06)**	8.26 (±0.06)
L. Thumb	14.23 (±0.18)	11.63 (±0.13)	9.70 (±0.06)	6.62 (±0.05)	6.46 (±0.05)	**6.15 (±0.05)**	6.63 (±0.05)
R. Thumb	13.84 (±0.18)	11.96 (±0.09)	10.36 (±0.06)	7.42 (±0.05)	7.29 (±0.05)	**6.99 (±0.05)**	7.47 (±0.06)
Average	12.26 (±1.45)	10.54 (±1.33)	8.92 (±1.16)	6.23 (±0.85)	6.04 (±0.86)	**5.78 (±0.83)**	6.25 (±0.85)

50%↑ Improvement

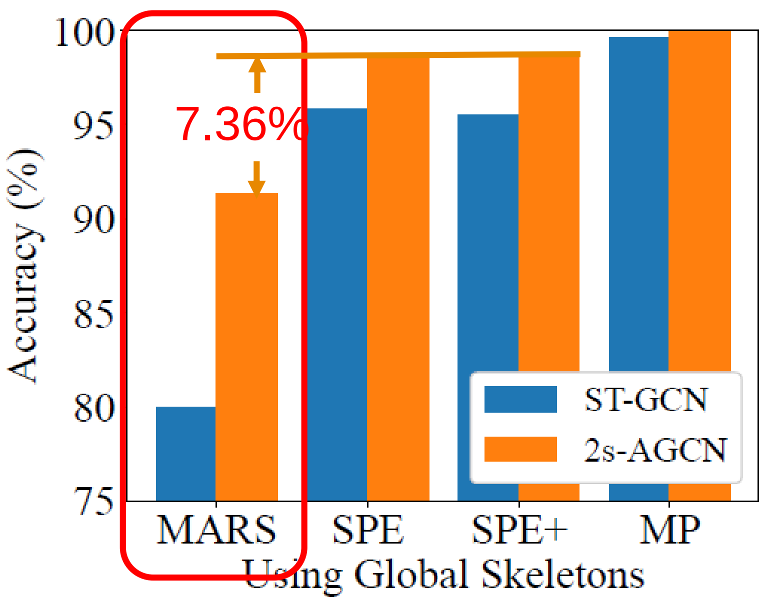


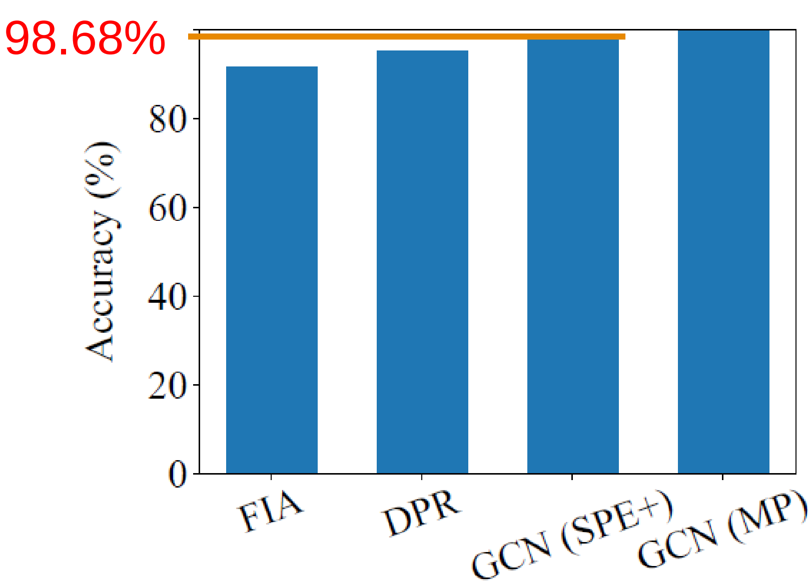
GCN – Experiment Setup

- GCN algorithms
  - ST-GCN
  - 2S-AGCN
- Input skeletons
  - MARS
  - SPE
  - SPE+
  - Mediapipe (MP)
- Baseline: FIA, DPR

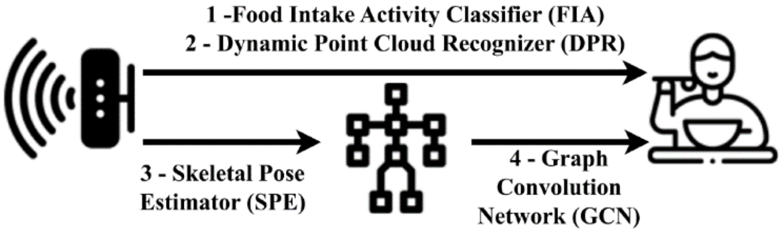
GCN Beats The End-to-End Solutions

- The quality of estimated skeleton affects the performance
- SPE achieved accuracies of 95.79% and 98.57% in the respective models, while SPE+ achieved 95.48% and 98.68% in ST-GCN & 2S-AGCN algorithms





Evaluation Summary



- FIA outperformed the SOTA RadHAR classifier, reached over 90% for both datasets
- DPR saves the memory but also has better performance than FIA
- SPE/SPE+ is the SOTA mmWave skeleton estimator
- SPE's skeleton with GCN achieves the best performance, 99% accuracy in "how" dataset

Outline

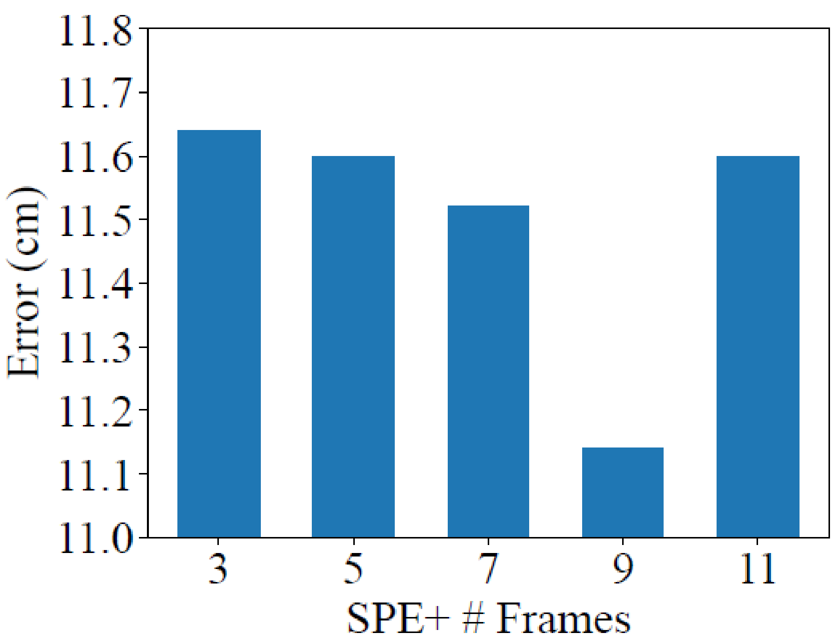
- Introduction
- Goal & Challenges
- Related Work
- Proposed Solutions
- Dataset
- Global Model Evaluations
- Leave-one-out Model Evaluations
- Conclusion & Future Work

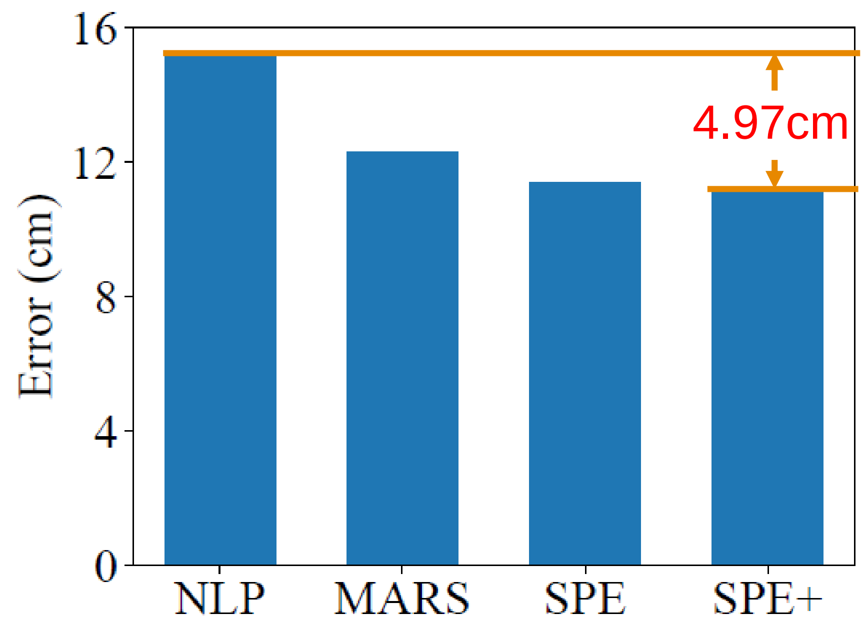
Leave-One-Out Model Setup

- 23 – 1 train/test split in our case
- Data from each subject will **ONLY** appear in the training **OR** testing sets
- Simulate the scenario in the subject is a new user of the system
- Cross-validation is performed

SPE+ is The Best Algorithm in Leave-One-Out Setup

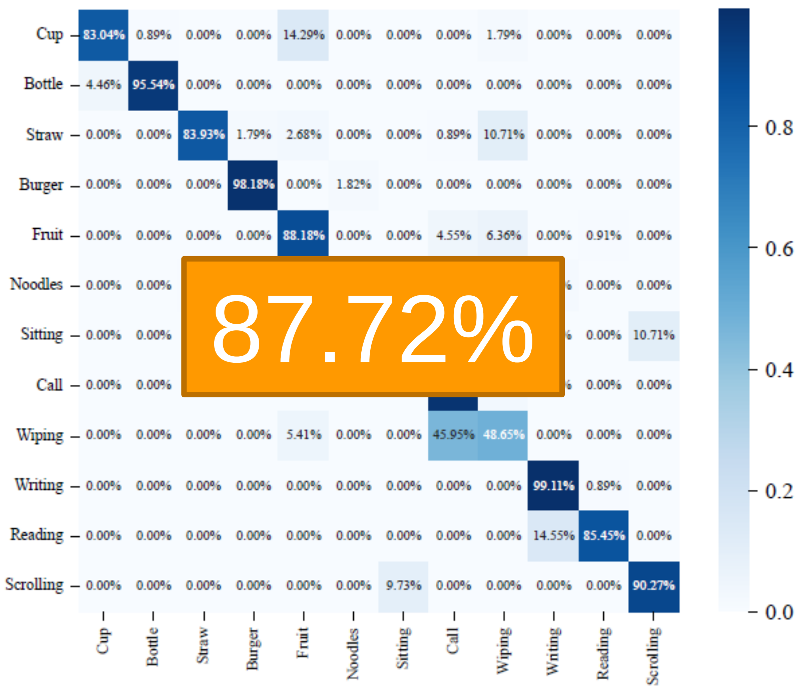
- SPE/SPE+ has the best performance with the 9 frames temporal aggregation setup
- SPE+ has the best performance of 11.14 cm, while NLP, MARS, and SPE get 15.11, 12.29 , and 11.38 cm

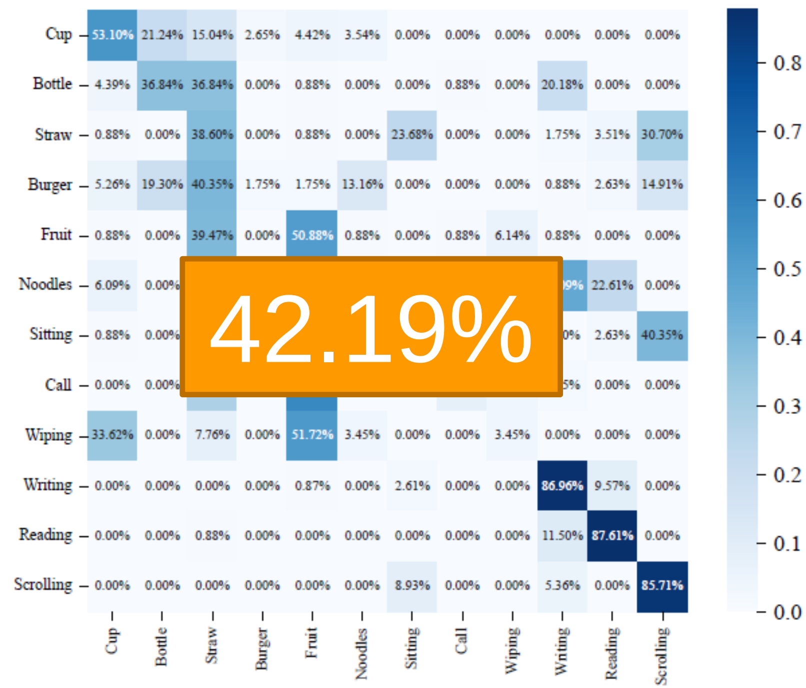




DPR Reaches Good Performances in Leave-One-Out Setup

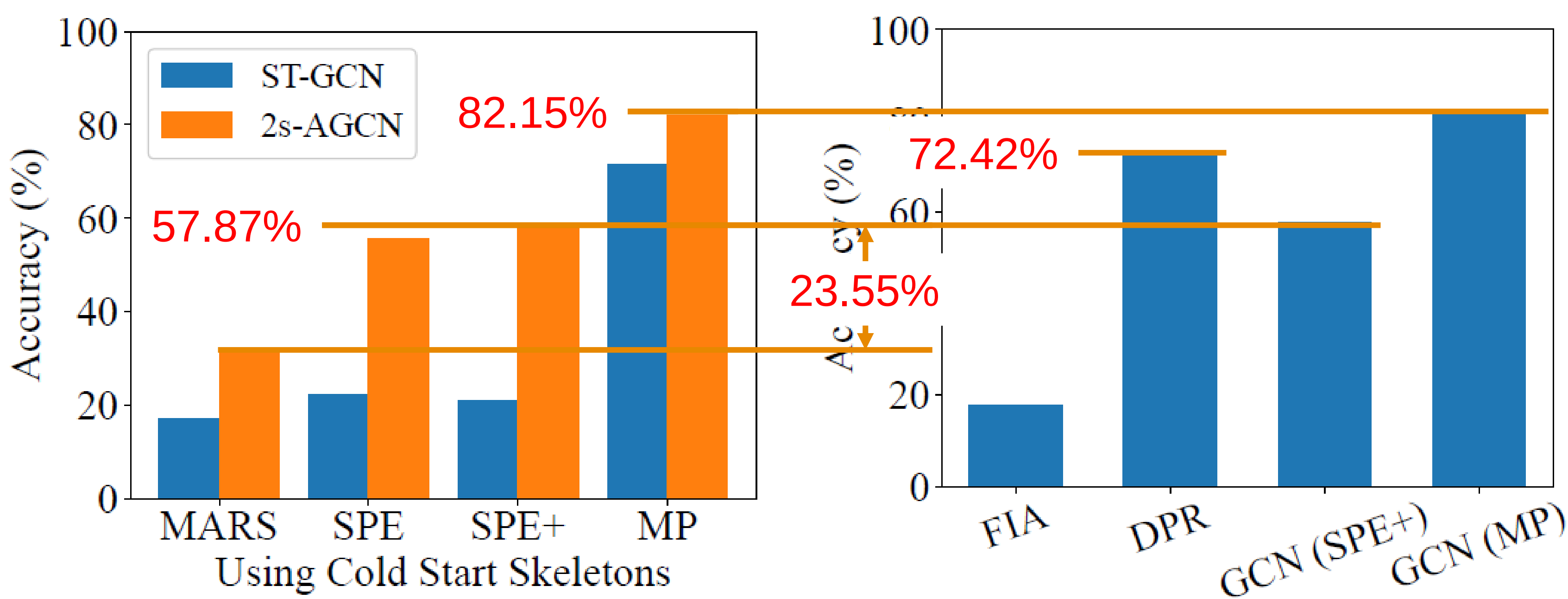
- DPR achieves an accuracy of 72.74%
- Performance varied by subject



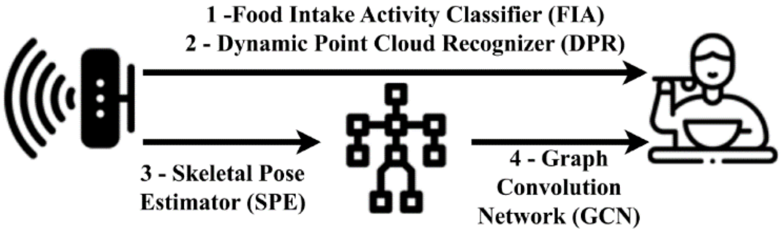


GCN Classifiers Suffered From Long Error Distance of Estimated Skeleton

- Quality of the estimated skeletons affects the performance critically
- GCN has the potential to outperform DPR if the skeleton quality was improved



Evaluation Summary

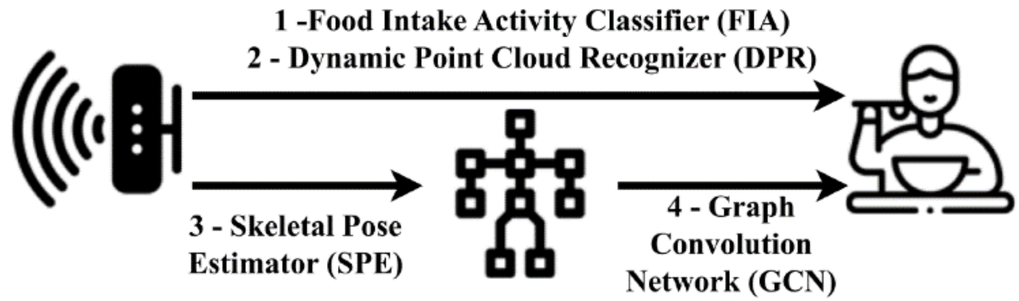


- SPE/SPE+ is the SOTA mmWave skeleton estimator
- DPR has the best performance in leave-one-out setup, with 72.42% accuracy
- GCN’s performance is highly affected by the quality of the skeletons

Outline

- Introduction
- Goal & Challenges
- Related Work
- Proposed Solutions
- Dataset
- Global Model Evaluations
- Leave-one-out Model Evaluations
- Conclusion & Future Work

Conclusions

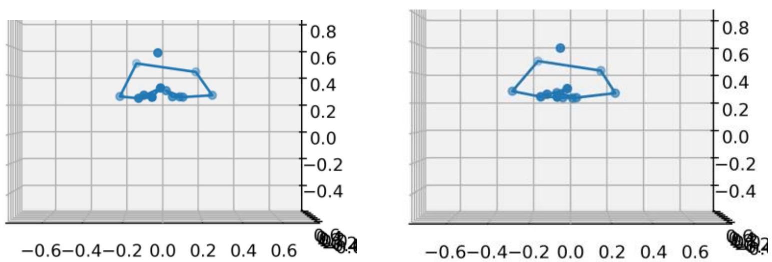


- 4 algorithms and a dataset is proposed to recognize food intake activities with mmWave radar point cloud
- SPE/SPE+ reaches the SOTA performance in fine-grained skeleton estimation
- SPE+ & 2s-AGCN classifier reaches 99% accuracy in global setup
- DPR achieved 72.46% accuracy in leave-one-out setup

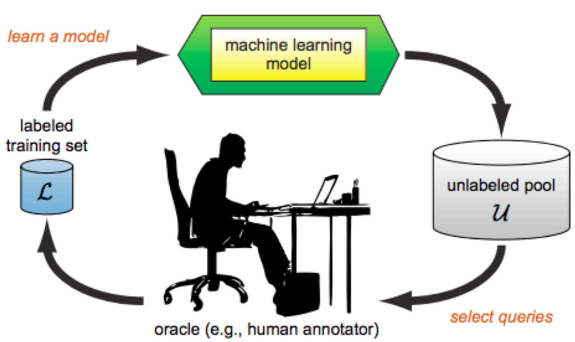
Future Work

The Leave-one-out Setup Issue

- Refinement of the skeleton models

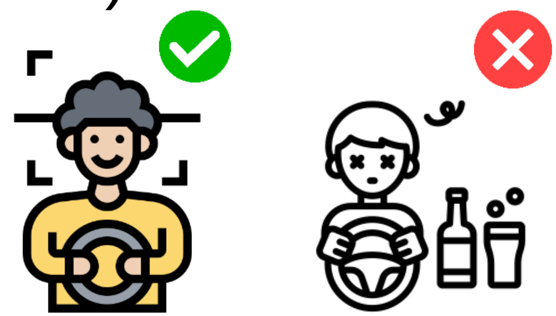


- Transfer learning as personalization model

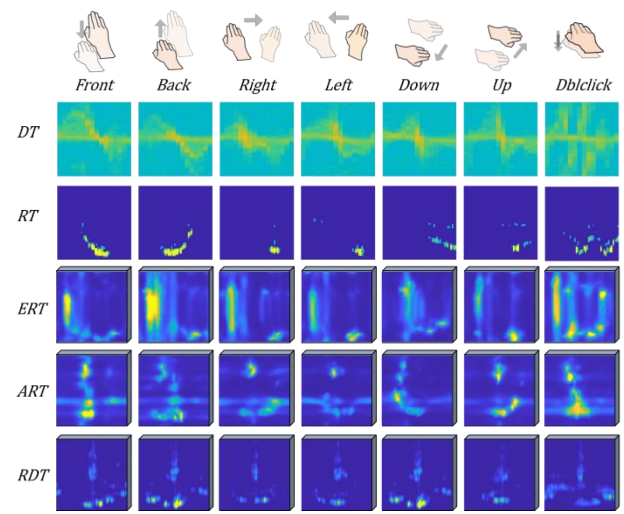


Other activity recognition problems

- Driver Monitoring System (DMS)



- Gesture Recognition



[1] ODSC. Active Learning: Your Model's New Personal Trainer
[2] Q.Chen. MIMOGR:MIMO millimeter wave radar multi-feature dataset for gesture recognition.

Thank you for listening!

Thanks for the help of Prof. Hsu, Prof. Shervin Shirmohammadi, Hsin-che Chiang, Yuanjie Chen, and all lab mates.

Publications:

Y.-H. Wu, Y. Chen, S. Shirmohammadi, and C.-H. Hsu. Ai-assisted food intake activity recognition using 3D mmwave radars. In Proc. of the ACM International Workshop on Multimedia Assisted Dietary Management (MADiMa), pages 81–89, 2022.

Y.-H. Wu, H.-C. Chiang, S. Shirmohammadi, and C.-H. Hsu. A dataset of food intake activities using sensors with heterogeneous privacy sensitivity levels. In Proc. of ACM Multimedia Systems, pages 416–422, 2023.

H.-C. Chiang, Y.-H. Wu, S. Shirmohammadi, and C.-H. Hsu. Memory-Efficient High-Accuracy Food Intake Activity Recognition with 3D mmWave Radars. In Proc. of ACM International Workshop on Multimedia Assisted Dietary Management (MADiMa). 2023

Q&A