LIGHTWEIGHT HUMAN ACTIVITY RECOGNITION FOR AMBIENT ASSISTED LIVING

Mohamad Reza Shahabian Alashti,
Mohammad Hossein Bamorovat Abadi,
Patrick Holthaus,
Catherine Menon,
And
Farshid Amirabdollahian

Robotics Research Group,
School of Engineering and Computer Science
University of Hertfordshire,
Hatfield, United Kingdom

University of Hertfordshire UH

IARIA

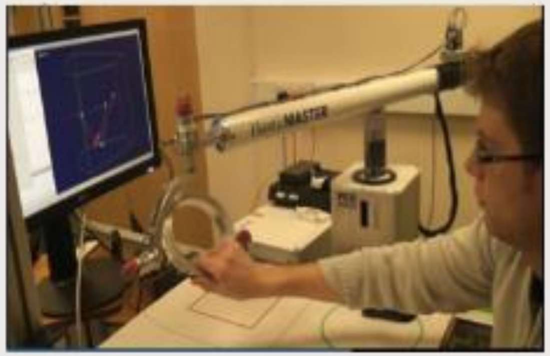
Email: {m.r.shahabian , m.bamorovat, p.holthaus, c.menon, f.amirabdollahian2}@herts.ac.uk

PRESENTER'S BIOGRAPHY

Mohamad Reza
Shahabian Alashti

I am a researcher with a passion for robotics, machine learning, and computer vision. I am grateful for the opportunity to pursue my PhD in Computer Science at the University of Hertfordshire, where I am currently working on skeleton-based human activity recognition using multiple cameras. Work in this area has the potential to revolutionize the way we interact with technology and could lead to new innovations in fields such as healthcare, sports, and entertainment. I am proud of my accomplishments; I remain committed to continued learning and growth in my field. I believe that my work has the potential to make a positive impact on society, and I am excited to continue exploring new applications for robotics and machine learning.

Robots in UH **Robot House** and Robotics Research Group



INTRODUCTION

- Multi-View Human Activity Recognition (MV-HAR) is an extension of traditional HAR that uses multiple views or perspectives to improve recognition performance.
- Indoor environments can be captured using multiple cameras or sensors to achieve a more robust and accurate recognition.
- A lightweight pipeline is important for real-time and resource-constrained applications, such as mobile devices, where computational efficiency and low power consumption are key requirements.
- Using a lightweight MV-HAR pipeline can provide a complete and accurate understanding of activities for assistive living systems.

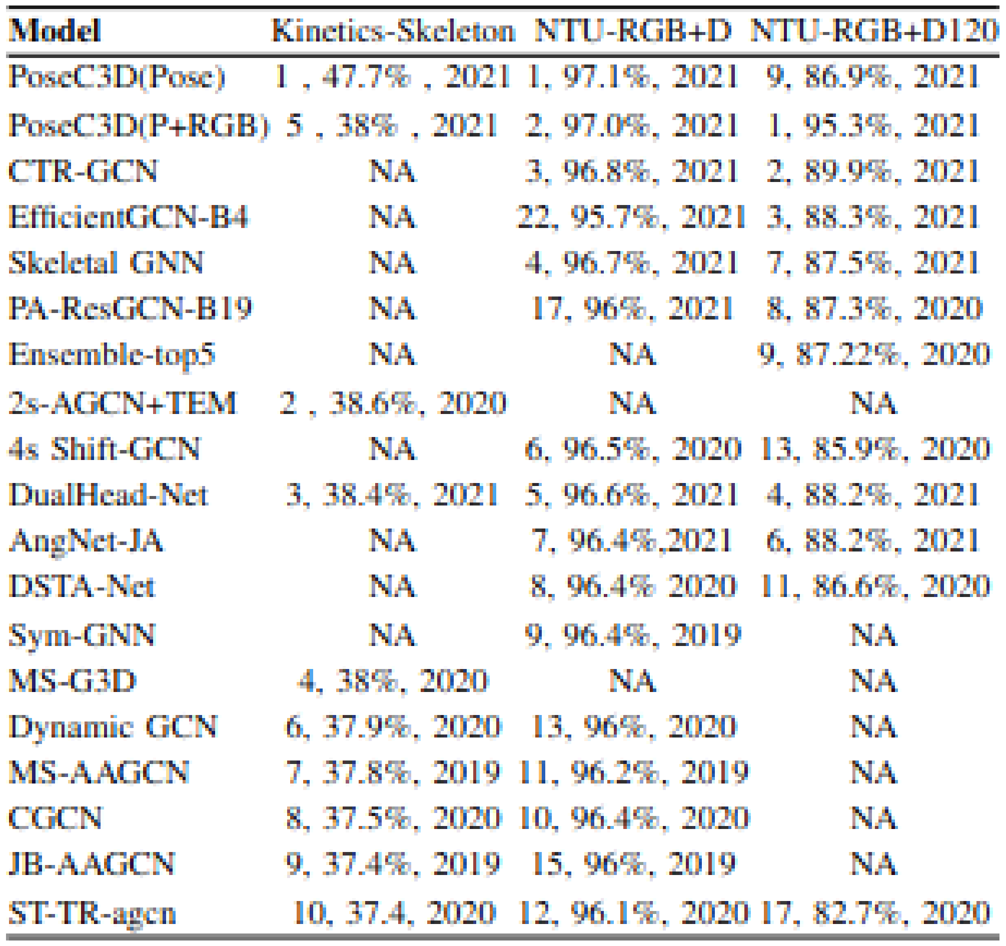
WE PRESENT TWO MAIN CONTRIBUTIONS

- **Development of a lightweight HAR pipeline**
  - Data sampling, input data type, and representation and classification.
- **Comparison of camera views**
  - Model execution in support of an experiment to find the performance of individual views and their combination for M-LeNet and ViT.

BACKGROUND

- The number of skeleton-based HAR methods is increasing, but there is still room for improvement
- Dataset details directly affect the accuracy of machine learning models
- The same model may not perform as well in a different dataset
- highlighting the need for comprehensive benchmarks to evaluate HAR algorithms
- Dataset specialization, based on theme, activity, task, and subject, can be used to address this challenge
- Our work aims to apply HAR in the AAL (Assistive Ambient Living) context using a skeleton-based and multi-view dataset.

TABLE I. Results Of Skeleton-Based HAR Leader Board In Three Datasets



Model	Kinetics-Skeleton	NTU-RGB+D	NTU-RGB+D120
PoseC3D(Pose)	1 , 47.7% , 2021	1, 97.1%, 2021	9, 86.9%, 2021
PoseC3D(P+RGB)	5 , 38% , 2021	2, 97.0%, 2021	1, 95.3%, 2021
CTR-GCN	NA	3, 96.8%, 2021	2, 89.9%, 2021
EfficientGCN-B4	NA	22, 95.7%, 2021	3, 88.3%, 2021
Skeletal GNN	NA	4, 96.7%, 2021	7, 87.5%, 2021
PA-ResGCN-B19	NA	17, 96%, 2021	8, 87.3%, 2020
Ensemble-top5	NA	NA	9, 87.22%, 2020
2s-AGCN+TEM	2 , 38.6%, 2020	NA	NA
4s Shift-GCN	NA	6, 96.5%, 2020	13, 85.9%, 2020
DualHead-Net	3, 38.4%, 2021	5, 96.6%, 2021	4, 88.2%, 2021
AngNet-JA	NA	7, 96.4%,2021	6, 88.2%, 2021
DSTA-Net	NA	8, 96.4% 2020	11, 86.6%, 2020
Sym-GNN	NA	9, 96.4%, 2019	NA
MS-G3D	4, 38%, 2020	NA	NA
Dynamic GCN	6, 37.9%, 2020	13, 96%, 2020	NA
MS-AAGCN	7, 37.8%, 2019	11, 96.2%, 2019	NA
CGCN	8, 37.5%, 2020	10, 96.4%, 2020	NA
JB-AAGCN	9, 37.4%, 2019	15, 96%, 2019	NA
ST-TR-agcn	10, 37.4, 2020	12, 96.1%, 2020	17, 82.7%, 2020

Three values in datasets' row define the *Rank* , *Accuracy*, and *Year* of publication respectively.

BACKGROUND

LIGHTWEIGHT APPROACH FOR MV-HAR

- MV-HAR focuses on using multiple perspectives of an activity to improve recognition performance
- Deep neural networks, convolutional neural networks, recurrent neural networks, and attention-based models have been proposed to enhance recognition accuracy
- Developing a comprehensive and real-world activity recognition system is demanding due to the extensive data and processing power required by some deep learning approaches
- A lightweight machine learning approach is essential for long-term deployment in assistive living scenarios
- Low computational cost, fewer training parameters, and efficient algorithms enable the system to be more practical
- Some high-accuracy single-view models such as PoseC3D and 2s-AGCN+TEM have 2m to 8m parameters and 6.94m parameters, respectively
- Models in MV-HAR with multiple views could have significantly more parameters

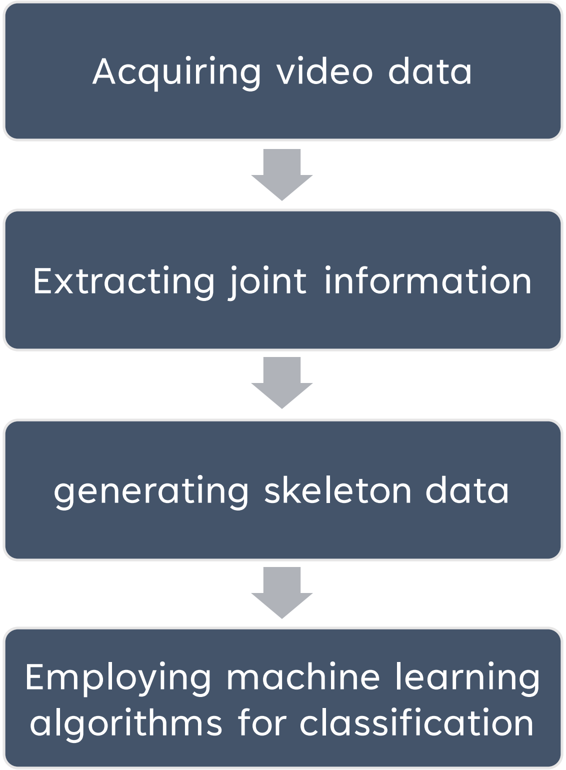
LIGHTWEIGHT HAR PIPELINE

- Development of a lightweight HAR pipeline includes **data sampling**, i**nput data type, and representation** and **classification**.

- Lightweight pipelines are important for real-time and resource-constrained applications, such as those on mobile devices, where computational efficiency and low power consumption are key requirements.

- A lightweight MV-HAR pipeline can enable more widespread deployment of activity recognition technology in smart homes or smart cities.

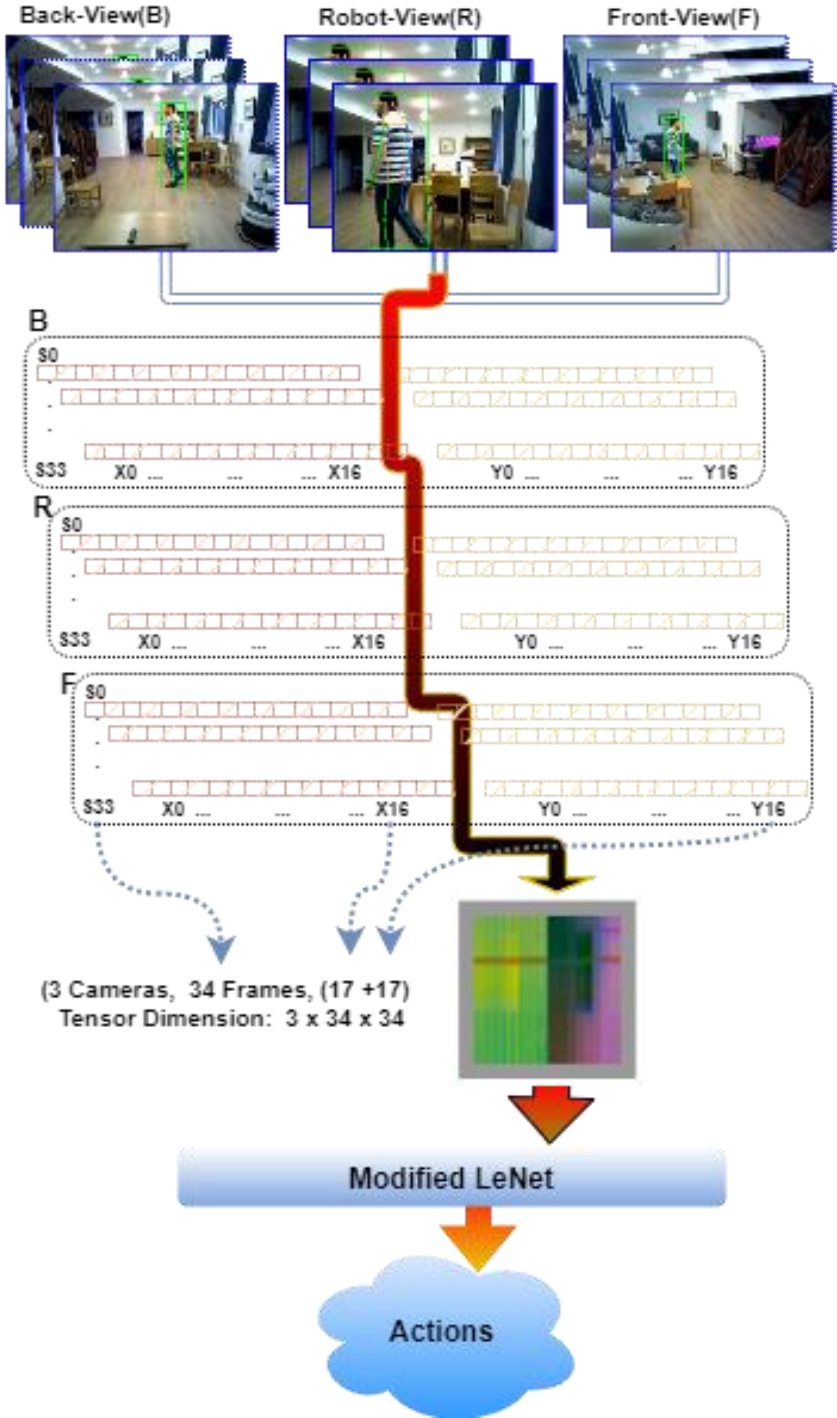
- The main goal is to develop a lightweight machine learning approach for real-time and resource-constrained applications such as robots in assistive living scenarios


Acquiring video data
Extracting joint information
generating skeleton data
Employing machine learning algorithms for classification

MV-HAR PIPELINE

Fourteen daily actions

[walking, bending, sitting down, standing up, cleaning, reaching, drinking, opening can, closing can, carrying object, lifting object, putting down object, stairs climbing up, stairs climbing down]



(3 Cameras, 34 Frames, (17 +17)
Tensor Dimension: 3 x 34 x 34

Modified LeNet

Actions

MV-HAR PIPELINE

MODIFIED LENET MODEL (M-LENET)

- The base model used in this experiment for CNN-based machine learning model is LeNet, a simple convolution model for image representation.
- Two convolution layers are applied in this model, which we test by two different configurations, 10 and 20 channels for the low parameter and 20 and 40 channels for the high parameter configuration.
- The difference between the original LeNet and this modified version is the number of convolution layers and the kernel size.
- Two dropout layers have also been added to avoid over-fitting.

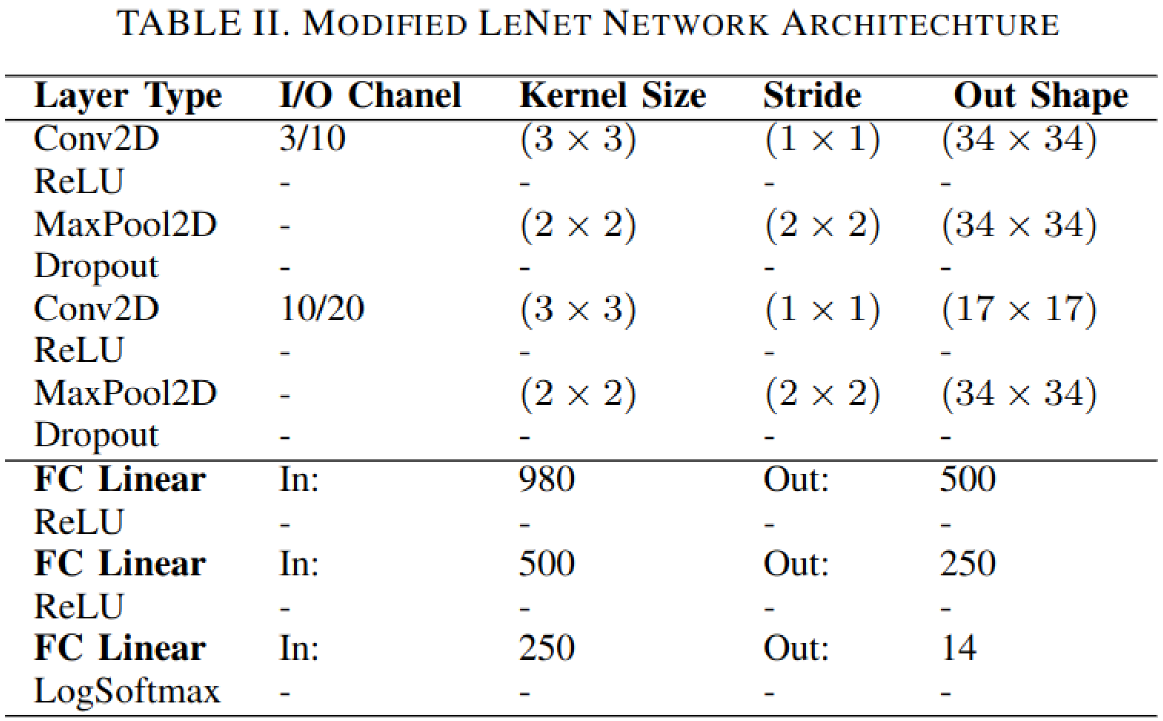


TABLE II. Modified LeNet Network Architechture

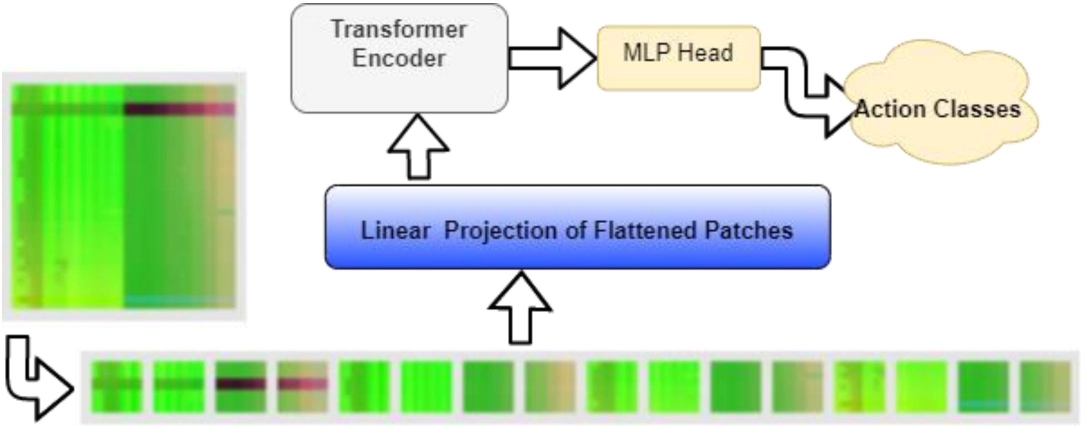
Layer Type	**I/O Chanel**	**Kernel Size**	**Stride**	**Out Shape**
Conv2D	3/10	(3×3)	(1×1)	(34×34)
ReLU	-	-	-	-
MaxPool2D	-	(2×2)	(2×2)	(34×34)
Dropout	-	-	-	-
Conv2D	10/20	(3×3)	(1×1)	(17×17)
ReLU	-	-	-	-
MaxPool2D	-	(2×2)	(2×2)	(34×34)
Dropout	-	-	-	-
FC Linear	In:	980	Out:	500
ReLU	-	-	-	-
FC Linear	In:	500	Out:	250
ReLU	-	-	-	-
FC Linear	In:	250	Out:	14
LogSoftmax	-	-	-	-

MV-HAR PIPELINE

VISION TRANSFORMERS (VIT) ARCHITECTURE

- Popular transformers-based image classification method
- Utilizes a self-attention mechanism to efficiently learn the relationships between different parts of an image.
- ViT achieved state-of-the-art results on the ImageNet dataset with a **top-1** accuracy of **90.2%** and **top-5** accuracy of **98.5%**, outperforming previous state-of-the-art methods such as **ResNet** and **EfficientNet**.

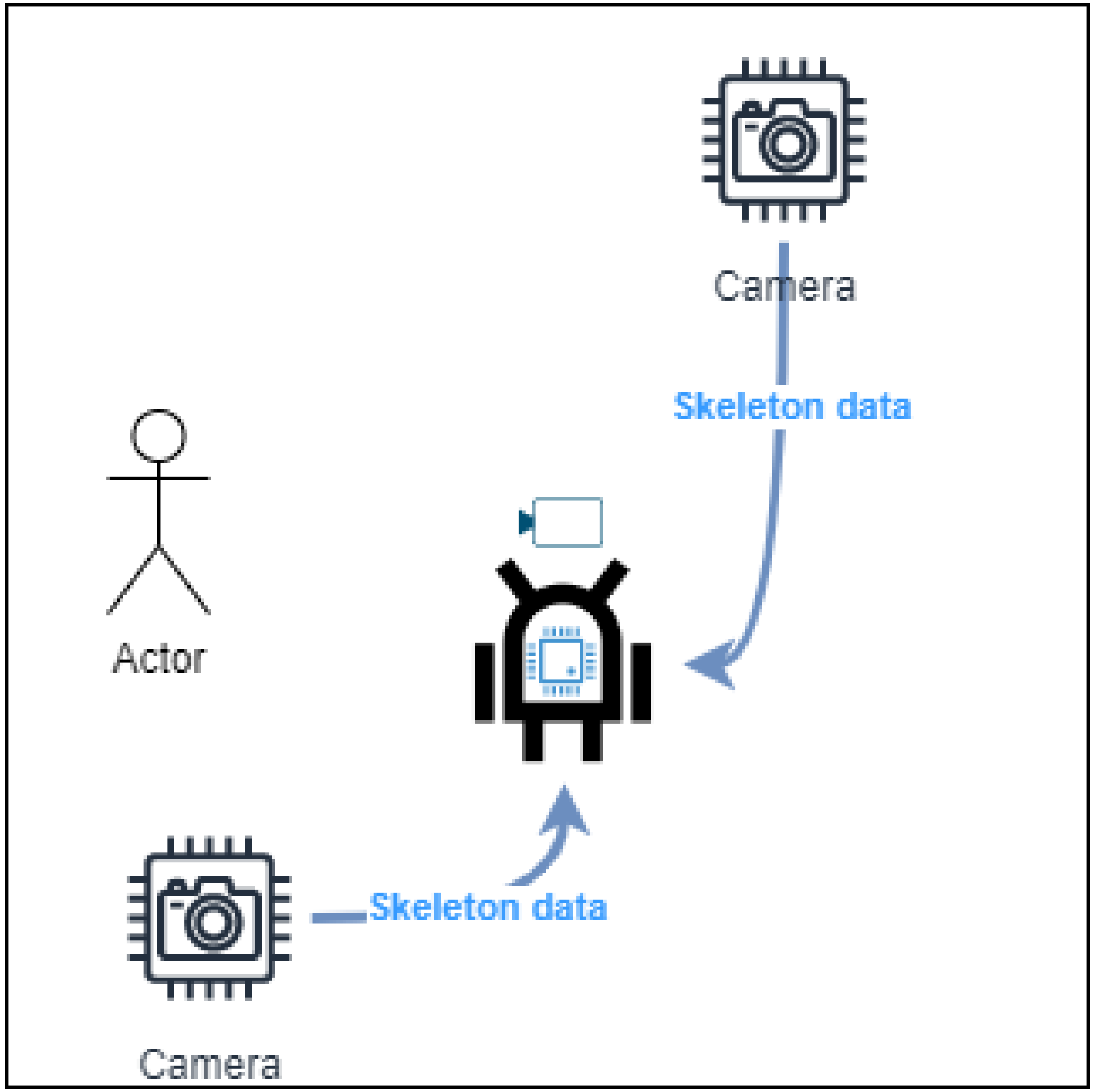
- Each input picture is divided into patches of sub-images
- Then by applying the positional encoding, the model is trained
- Each patch is considered a word and projected to the feature space



MV-HAR PIPELINE

DECENTRALIZED STRUCTURE

- Multiple cameras with separate processors offers numerous advantages
- Extracting and transmitting only the crucial skeleton information reduces the robot's computational load, making it more efficient and responsive in providing assistance.
- The use of multiple cameras can enhance the accuracy of the interaction, as the robot can take inputs from different angles into account.
- This leads to a more human-like interaction, which is crucial in assistive settings where the goal is to create a seamless and intuitive experience.
- It makes the assistive robot even more efficient in providing aid.
- Overall, this approach significantly enhances the capabilities of assistive robots and provides a better experience for those in need of assistance.



Comparison of M-LeNet and ViT Models

on RHM-HAR-SK Dataset

RESULTS

- Additional view can enhance the robot view
- Lightweight model contribute lower parameters
- Fusion of multiple views with the same params
- Removing low accuracy positions(previous work) doesn’t affect accuracy but has lower params
- Despite the simple structure of M-LeNet, it is competitive compared to the ViT

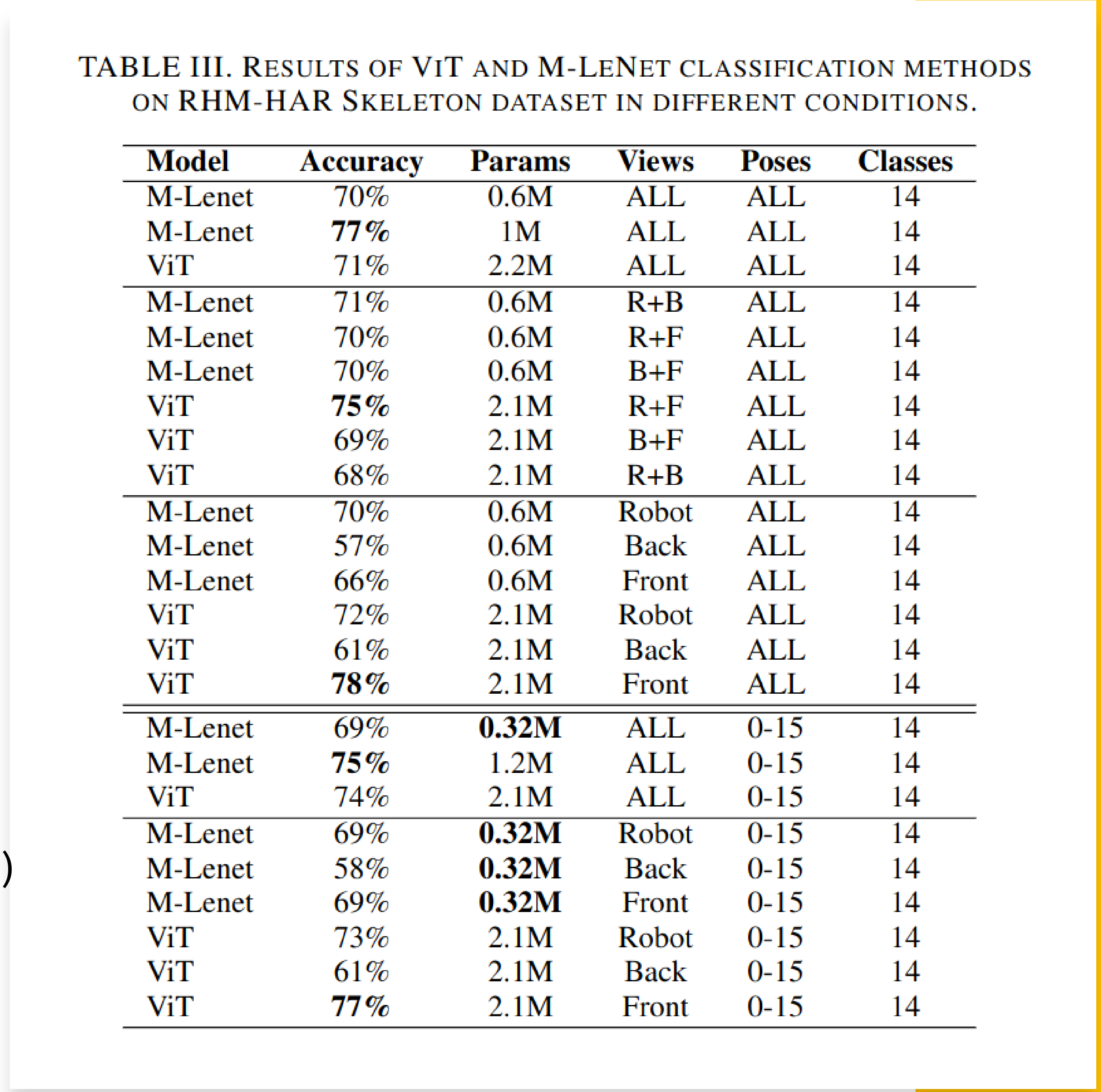


TABLE III. RESULTS OF VIT AND M-LENET CLASSIFICATION METHODS ON RHM-HAR SKELETON DATASET IN DIFFERENT CONDITIONS.

Model	**Accuracy**	**Params**	**Views**	**Poses**	**Classes**
M-Lenet	70%	0.6M	ALL	ALL	14
M-Lenet	**77%**	1M	ALL	ALL	14
ViT	71%	2.2M	ALL	ALL	14
M-Lenet	71%	0.6M	R+B	ALL	14
M-Lenet	70%	0.6M	R+F	ALL	14
M-Lenet	70%	0.6M	B+F	ALL	14
ViT	**75%**	2.1M	R+F	ALL	14
ViT	69%	2.1M	B+F	ALL	14
ViT	68%	2.1M	R+B	ALL	14
M-Lenet	70%	0.6M	Robot	ALL	14
M-Lenet	57%	0.6M	Back	ALL	14
M-Lenet	66%	0.6M	Front	ALL	14
ViT	72%	2.1M	Robot	ALL	14
ViT	61%	2.1M	Back	ALL	14
ViT	**78%**	2.1M	Front	ALL	14
M-Lenet	69%	**0.32M**	ALL	0-15	14
M-Lenet	**75%**	1.2M	ALL	0-15	14
ViT	74%	2.1M	ALL	0-15	14
M-Lenet	69%	**0.32M**	Robot	0-15	14
M-Lenet	58%	**0.32M**	Back	0-15	14
M-Lenet	69%	**0.32M**	Front	0-15	14
ViT	73%	2.1M	Robot	0-15	14
ViT	61%	2.1M	Back	0-15	14
ViT	**77%**	2.1M	Front	0-15	14

CONCLUSION

- Proposed a lightweight multi-view skeleton-based human activity recognition (HAR) method for enhancing ambient assisted living scenarios.
- The pipeline combines the advantages of both multi-view and skeleton-based activity recognition by fusing information from multiple RGB cameras to enhance the activity perception of the AAL system.
- Utilized a modified LeNet classification model and Vision Transformer for the classification task.
- Performance assessment found that combining camera views can improve recognition accuracy.
- The proposed pipeline presents a more efficient and scalable solution for ambient assisted living systems, thus providing a potential for improving the safety, comfort and quality of life for AAL users.

THANK YOU