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

***Abstract*—Ambient Assisted Living (AAL) systems aim to improve the safety, comfort, and quality of life for the populations with specific attention given to prolonging personal independence during later stages of life. Human Activity Recognition (HAR) plays a crucial role in enabling AAL systems to recognise and understand human actions. Multi-view human activity recognition (MV-HAR) techniques are particularly useful for AAL systems as they can use information from multiple sensors to capture different perspectives of human activities and can help to improve the robustness and accuracy of activity recognition. In this work, we propose a lightweight activity recognition pipeline that utilizes skeleton data from multiple perspectives with the objective of enhancing an assistive robot's perception of human activity. The pipeline includes data sampling, spatial temporal data transformation, and representation and classification methods. This work contrasts a modified classic LeNet classification model (M-LeNet) versus a Vision Transformer (ViT) in detecting and classifying human activities. Both methods are evaluated using a multi-perspective dataset of human activities in the home (RHM-HAR-SK). Our results indicate that combining camera views can improve recognition accuracy. Furthermore, our pipeline provides an efficient and scalable solution in the AAL context, where bandwidth and computing resources are often limited.**

***Keywords*—classification, Multi-view, Skeleton-based, Activity Recognition Pipeline, Assistive Robot**

I. Introduction

Multi-View Human Activity Recognition (MV-HAR) is an extension of traditional HAR in which multiple views or perspectives of activity are used to improve recognition performance. This is thought to be beneficial due to ability to provide a clear and undisturbed view of a dynamic activity. In indoor environments, this can be achieved by using multiple cameras or sensors to capture different views of the same activity and then fusing the information provided by different views to achieve a more robust and accurate recognition.

The process of MV-HAR typically involves capturing video or sensor data, pre-processing the data to extract features, and then using machine learning algorithms to classify the activities. A lightweight pipeline is important for real-time and resource-constrained applications, such as those on mobile devices like robots, where computational efficiency and low power consumption are key requirements. Additionally, lightweight pipelines can enable more widespread deployment of activity recognition technology, such as in smart homes or smart cities, where large numbers of cameras or sensors need to be integrated.

For assistive living systems, using a lightweight MV-HAR pipeline can provide a complete and accurate understanding of the activities performed by residents, including older adults or people with disabilities. This can enable the development of more effective and personalized support services, such as fall detection and home security, and also supports principles of prevention and pro-active care.

In this work, we modify a classic LeNet [11] classification model, termed as M-LeNet for the HAR task. To contrast, we also use the Vision Transformer [7] (ViT) for the classification task, and compare the results between both models. The rationale behind selecting the LeNet and Vision Transformer (ViT) classifiers for the Human Activity Recognition (HAR) task stems from their distinctive characteristics in terms of architecture and design. Specifically, LeNet is a widely used and relatively simplistic Convolutional Neural Network (CNN) in image classification tasks, whereas ViT is a more advanced transformer-based model that has gained popularity in recent years owing to its ability to process images without relying on traditional convolutional layers. Both classifiers were chosen due to their comparable number of training parameters, thus allowing for a fair comparison between the two models. Besides, several parts of the HAR pipeline like input spatial temporal data transformation, data sampling, and representation and classification methods have been modified.

With this work, we therefore present:

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.

To provide context for our approach, we first review related works in Sec. II that have been applied to three popular HAR datasets, and discuss the importance of multi-view datasets for assistive robots scenarios. In Sec. III, we present a new lightweight multi-view pipeline and provide a detailed explanation of its structure. In Sec. IV, we evaluate the performance of different camera views in terms of accuracy and number of parameters of two different classifier models. Finally, we conclude our paper in Sec. V.

II. Related Work

A. Skeleton-based Methods

Based on the spatial and temporal nature of human activity, different methods have emerged. *Sequence models* like Recurrent Neural Network (RNN) [12] or Long Short Term Memory (LSTM) influence the sequentially of extracted human skeleton data as time series. *Convolutional Neural Network* (CNN) based models [14] have great potential in spatial information compared to RNN models. The other successful methods are *Graph Neural network* based (GNN) [15], [25] which represent spatial and temporal information by the human skeleton's natural topological graph structure. Spatio-Temporal Graph Convolutional Network (ST-GCN) [25] is the first model in this category that notices harmony in spatial and temporal data that allows for combining spatial structures with time-series while still benefiting from a convolutional neural network.

Besides, *transformer* models have been engaged in HAR tasks to gain competitive outcomes. They can be used to capture long-range dependencies between regions of an image, allowing the model to better express understand the relationships between objects and their context [7]. Some of them rely on modified GCN models [18], [26] and others [21] are purely transformer based.

B. Skeleton-based HAR Leader board analysis

The investigation of skeleton-based action recognition reveals that NTU-RGB+D [24], NTU-RGB+D 120 [13], and Kinetics-skeleton [10] datasets are trending nowadays. Table I illustrates these datasets' top-ranked skeletal model performances. The rank number, model's accuracy, and year of publication have been provided to show the diversity of ML models and their sometimes varying behavior in different datasets.

The *PoseC3D* [8] method has the highest accuracy in two datasets (Kinetics-Skeleton and NTU-RGB+D) and stands at rank nine in one other. In addition, a different variation of PoseC3D, RGB + Pose, has ranked five in kinetics skeleton and first and second rank in two others.

Considering the available number of Skeleton-based HAR models, NTU RGB+D has the highest with 85, followed by NTU RGB+D 120 with 38, and then Kinetics-skeleton with 18. In Table I, the top ten models in terms of accuracy in almost all datasets have been considered. Kinetics-skeleton is the base dataset for sorting the models ranks. Given that not all models are applied in all three datasets, comparable results are not always available. The total number of models is 21.

The range of accuracy is not the same in all datasets. The highest performance in Kinetics-Skeleton dataset belongs to PoseC3D (w.HRNet 2D skeleton) with 47.7%, following that by almost 9% difference, the 2s-AGCN+TEM [17] model accuracy is 38.36%. Ironically, the rest of the models' accuracy was distributed in 1%, from 38.4% for DualHead-Net [5], the 3rd rank, to 37.4% for ST-TR-agcn [18] , the 10th rank. However, the total accuracy range of ranks in two other

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.

datasets is like a uniform distribution. Low difference, 0.7% for NTU RGB+D and 3%, for NTU RGB+D 120. However, based on this evidence, it is unreliable to say that a method is superior by considering its rank in just one dataset. For example, EfficientGCN-B4 [22] model stands in the third stage on the leader board for NTU RGB+D 120 dataset, but its rank in NTU RGB+D is 22. Likewise, PoseC3D (w. HRNet 2D skeleton), which has outstanding results in the Kinetics-skeleton dataset, and the highest accuracy in NTU RGB+D, stands in stage nine in the leader board for NTU RGB+D 120 dataset. However, another variant of PoseC3D (RGB + Pose) has conspicuous accuracy in NTU RGB+D 120 (95.3%) dataset and high accuracy in two others.

On the other hand, models like CTR-GCN [6], Skeletal GNN [28], 2s-AGCN+TEM [17], DualHead-Net, AngNet-JA + BA + JBA + VJBA [19], MS-G3D [15], CGCN [27], has reasonable accuracy because they stand in top rank in two or all datasets, respectively.

To summarise, although the number of skeleton-based human activity recognition methods and their variation is increasing, there is still room for improvements for models to be applied in challenging datasets like Kinetics-Skeleton. The comparison reveals that dataset details have a direct effect on the ML model accuracy. For example, kinetic-skeleton data is collected from Youtube videos and includes uncontrolled environment, and the NTU RGB+D videos were captured in a controlled environment. Top accuracy for the kinetic is almost 50% fewer than others. Besides, this review illustrates that the same model may not perform as well in a different dataset.

On the other hand, developing a comprehensive and real-world activity recognition is demanding, particularly given the nature of some *Deep-Learning* (DL) approaches, which

require extensive data and significant processing power e.g. CPU and GPU nodes. This results in a lack of comprehensive benchmarks [1] for evaluating the performance of activity recognition algorithms. One approach to solve this problem is dataset specialization, in which elements such as theme, activity, task, and subject adhere to a specific idea. In this work, we aim to apply HAR in the AAL context using a skeleton-based and multi-view dataset.

C. Multi-view HAR

Recent research in MV-HAR with skeleton models in indoor environments has focused on developing methods that can effectively utilize the temporal and spatial information provided by skeleton data. Methods such as deep neural networks [4], convolutional neural networks [24], recurrent neural networks, and attention-based models [2] have been proposed to improve the robustness and accuracy of the recognition system.

In MV-HAR systems, a lightweight machine learning approach is essential for providing real-time and resource-constrained applications like robots. A low computational cost, fewer training parameters, and an efficient algorithm enable the system to be more practical for long-term deployment in assistive living scenarios. However, focusing on the number of training parameters of the existing skeleton-based models shows that many methods are not computationally effective. For example, considering some single-view high-accuracy models in the Table I, *PoseC3D* [8] in different variation has 2m to 8m parameters and 2s-AGCN+TEM [17] has 6.94m parameters. Expanding the comparison to the multi-view, this could indicate models with significantly more parameters.

III. Lightweight MV-HAR pipeline

The process of recognizing human activity via a skeleton-based multi-view approach typically encompasses the acquisition or loading of video data, the extraction of joint information, and the generation of skeleton data. Subsequently, a machine learning algorithm is employed to classify the recorded actions. The utilization of a lightweight pipeline in this context allows for the integration of cameras in robotic and AAL systems, enabling their effective operation in a variety of scenarios. As depicted in Figure 1, our proposed methodology for Multi-View Human Activity Recognition (MV-HAR) emphasizes the central concept of leveraging multiple camera viewpoints to enhance the recognition of activities via a lightweight pipeline. The pipeline started with data collection from different views, followed by pose extraction and preprocessing. Then, the prepared tensor file feeds the training model.

A. Input Data

In a parallel work, we have developed the RHM-HAR-SK [20] dataset on the top of a RGB dataset (RHM) [3]. This non-generative multi-view skeleton-based human activity recognition dataset includes 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]* captured in an indoor typical British house. A robot-view camera, two wall-mounted cameras (Front-view and Back-view), and an omnidirectional view (Omni-view) camera capture the activities synchronously. However, analysis of that dataset reveals that the Omni-view data has low accuracy in the skeleton-based method, and it has consequently been omitted.

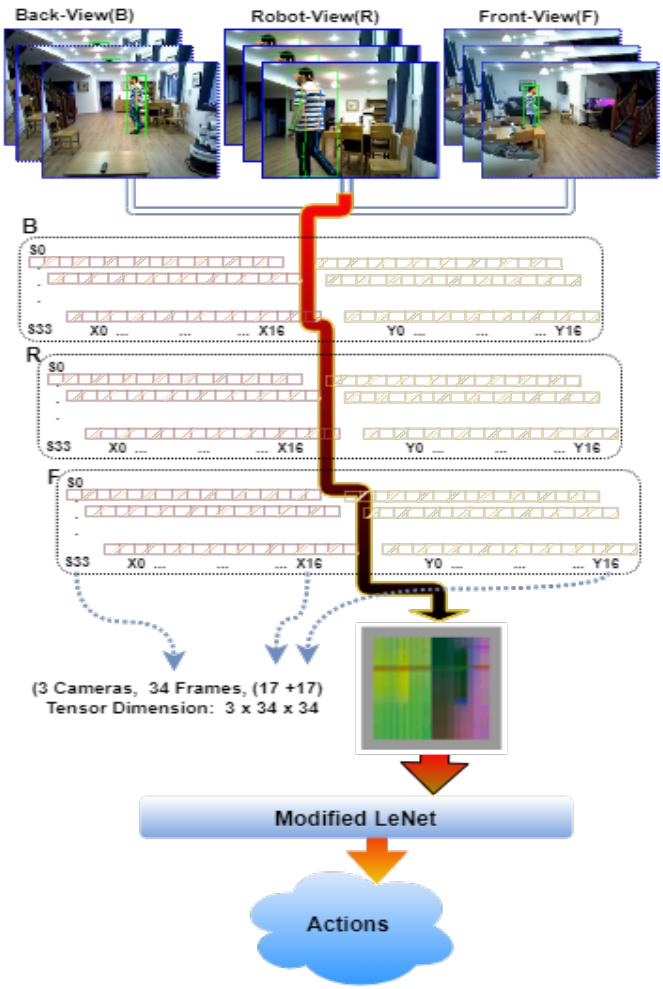


FIGURE 1. THE MV-HAR PIPELINE, AS DESCRIBED IN DETAIL IN SECTION III BEGINS WITH CAPTURING VIDEO FROM MULTIPLE VIEWS (III-A) AND THE EXTRACTION OF SKELETONS FROM MULTIPLE VIEWPOINTS (III-B, FOLLOWED BY THE CONVERSION OF EACH VIEWPOINT INTO A SPATIAL-TEMPORAL MATRIX (III-C). THESE MATRICES ARE SUBSEQUENTLY COMBINED INTO A SINGLE TENSOR FILE, WHICH IS FINALLY CLASSIFIED BY THE MODIFIED LENET MODEL (III-D).

B. Pose extraction

The utilization of RGB cameras is due to their simplicity, affordability, and accessibility in conjunction with a high-performance pose extraction method applied to RGB data, results in improved human body skeleton extraction. In RHM-HAR-SK dataset, a pretrained HRNet model as described in [23] is utilized to extract poses from videos. This model has been trained on the COCO keypoint detection dataset [9] and the MPII Human Pose dataset [16].

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	-	-	-	-

C. Preprocessing

Following the extraction of the skeleton data, the spatial and temporal input data was transformed into a $3 \times 34 \times 34$ tensor. In Figure 1 the process of finding a single person from three cameras to make the tensor file is shown. The first digit (3) refers to three cameras, and the 34×34 dimension refers to 34 frames of skeleton data, and two 17 columns. The first 17 columns belong to the *X* value and the second half to the *Y* value. Random sampling has been used to choose 34 frames for each video stream. The three-channel matrix is illustrated as an RGB image in Figure 1, with each camera view being mapped to the red, green, and blue channels.

Figure 2 illustrates three samples of two types input data, the RGB in 2b and grayscale in 2a. The former refers to three channels, each indicating a camera view and the latter a single-view camera. Each 2D image depicts skeletons data frames in an action. Subsequent to the extraction of skeletons from the video stream and preparing the 2D image, two general machine learning models were applied as outlined below.

D. The Modified LeNet model

The base model that has been used in this experiment for CNN-based machine learning model is LeNet [11]. This is a simple convolution model for image representation that we have modified as follows to use as the skeleton-based action classification. Table II illustrates the structure of the Modified CNN model. 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 (reduced from 3 to 2) and the kernel size (reduced from 5 to 3). Two dropout layers have also been added to avoid over-fitting. One more fully connected layer was also added to increase the learning parameters.

E. Vision Transformers (ViT) Architecture

The other classification model used is the ViT [7]. In the ViT architecture 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. In Figure 3 a random input data with its patches and the ViT classification architecture is shown. The process of preparing the input data for the ViT and M-LeNet is the same.

F. Decentralised structure

Implementing multiple cameras with separate processors in human-robot interactions 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. Figure 4 illustrates the proposed concept of decentralized structure of the multi-view camera with robotic agent. Two individual cameras refer the front-view and back-view in our experiment. The mobile robot with a camera is following the human to recognise its activities.

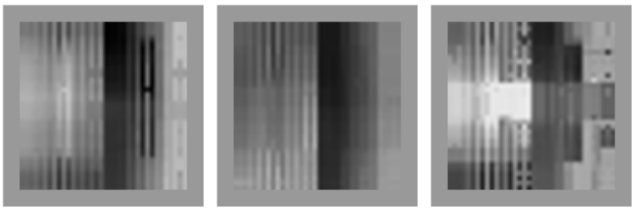
The use of multiple cameras can enhance the accuracy of the interaction, as the robot can take inputs from different angles into account. This leads to a more human-like interaction, which is crucial in assistive settings where the goal is to create a seamless and intuitive experience, making the assistive 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.

IV. RESULTS

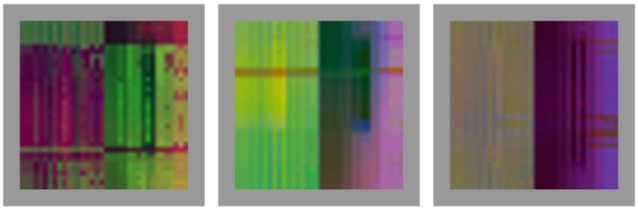
This section compares the results of the *M-LeNet* and *ViT* classification models in different conditions. Table III shows the comparison results of parameters like *accuracy*, number of total *trainable parameters*, different *camera views*, number of skeleton *positions*, and *classes*.

These two models are applied to the RHM-HAR-SK dataset, including a synchronized three-view video stream. Table III show the results of the models trained on full views, for all 14 classes, and all poses in the upper section and the lower section show the same comparison with excluded ankle poses (marked with 0-15 on poses column). The results show that the overall accuracy is between 69 and 77 percent for all views and 57 to 78 percent for single and double views. Among them, the comparison of models with all poses and removed ankles shows that for the ViT model, the accuracy moderately increased by 3% in all views and remains the same in a single view. In contrast, the M-LeNet model decreased by about 2% in both high and low parameters models. The difference between these M-LeNet classifiers is the number of parameters in linear layers, which one is double compared to the other. The high parameter M-LeNet has higher accuracy.

In the last part of the lower and upper section, the details of *single view* training models are shown. Interestingly, the ViT model results follow the missed poses statistics in the RHM-HAR-SK dataset [20], in which the front and robot views have fewer missed poses, and the highest accuracy among all views, and the back view is less accurate with more missed poses. Moreover, the front-view accuracy is 78% in ViT model, and robot-view accuracy is 70% in M-LeNet. The double-view combination results are shown in the second part of the upper section. Combinations of Front (F), Back (B)



(A) THREE SAMPLES OF SINGLE VIEW OF BENDING ACTION.



(B) THREE SAMPLES OF COMBINED THREE VIEWS AS A RGB IMAGE OF BENDING ACTION

FIGURE 2. SYNCHRONIZED SKELETON OUTPUT FROM DIFFERENT VIEWS OF BENDING ACTION.

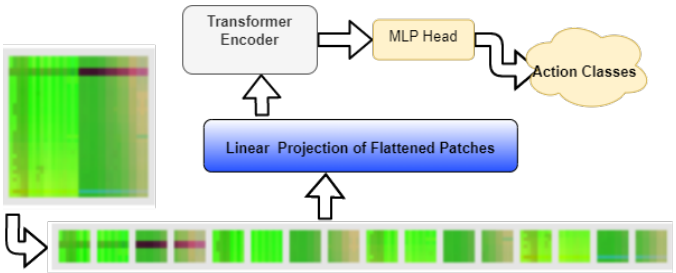


FIGURE 3. THE VIT CLASSIFICATION ARCHITECTURE APPLIED ON ONE OF THE RHM-HAR-SK DATASET'S SAMPLE [7]

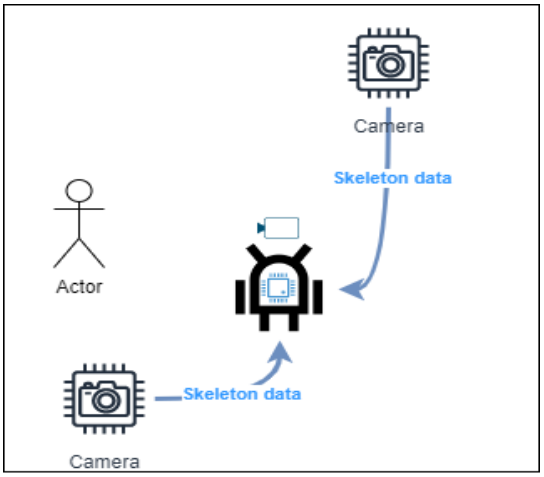


FIGURE 4. THE DECENTRALISED STRUCTURE OF MV-HAR WITH LOW COMPUTATIONAL COST IN THE ROBOT

and Robot (R) views are considered for assessing their impact on accuracy. The average accuracy of the double-view in the ViT models is higher than the lowest accuracy in the relevant single-view and less than the higher one, which means that the accuracy of the view with a lower value increased. For instance, the individual robot-view accuracy increased from 72% to 75% in combination with front-view and back-view increased from 61% to 69% when fused with front-view. For M-LeNet models, all single-view accuracy increased in the combination of double-views. The comparison of the upper section and lower one proves that removing low confidence joints like the ankle joints does not affect negatively even in

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

ViT all-views model accuracy increased by 3%. For M-LeNet and all single views, the accuracy fluctuated about 1%. An examination of the number of parameters in Table III illustrates that the M-LeNet model exhibits a significantly lower number of parameters in comparison to the ViT model. Furthermore, the results of removing poses with lower accuracy further contribute to the reduction in the model's parameters.

V. CONCLUSION

In this paper, we proposed a lightweight multi-view skeleton-based human activity recognition (HAR) method for enhancing ambient assisted living scenarios. The suggested 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. A modified LeNet classification model and Vision Transformer were utilized for the classification task. A performance assessment of the two models and their variations on a publicly available dataset found that combining camera views can improve recognition accuracy. Furthermore,

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. Our findings indicate that multiple recognition models, for example, *M-LeNet* and *ViT* could potentially be selected automatically based on information found in the scene, utilising the richness of captured data and information-theoretic modelling, which we plan to develop this further in our future work.

References

[1] Mohamad Reza Shahabian Alashti, Mohammad Hossein Bamorovat Abadi, Patrick Holthaus, Catherine Menon, and Farshid Amirabdollahian. Human activity recognition in robocup@ home: Inspiration from online benchmarks. *UKRAS21*, 2021. 3

[2] Yue Bai, Zhiqiang Tao, Lichen Wang, Sheng Li, Yu Yin, and Yun Fu. Collaborative attention mechanism for multi-view action recognition. *arXiv preprint arXiv:2009.06599*, 2020. 3

[3] Mohammad Bamorovat Abadi, Mohamad Reza Shahabian Alashti, Patrick Holthaus, Catherine Menon, and Farshid Amirabdollahian. Rhm: Robot house multi-view human activity recognition dataset. IARIA, March 2023. ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions, ACHI 2023 ; Conference date: 24-04-2023 Through 28-04-2023. 3

[4] Xiaobin Chang. *Deep Multi-View Learning for Visual Understanding*. PhD thesis, Queen Mary University of London, 2019. 3

[5] Tailin Chen, Desen Zhou, Jian Wang, Shidong Wang, Yu Guan, Xuming He, and Errui Ding. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4334–4342, 2021. 2

[6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 4, 5

[8] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 2, 3

[9] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1, 4

[12] Guy Lev, Gil Sadeh, Benjamin Klein, and Lior Wolf. Rnn fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*, pages 833–850. Springer, 2016. 2

[13] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 2

[14] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017. 2

[15] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020. 2

[16] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5137–5146, 2018. 3

[17] Yuya Obinata and Takuma Yamamoto. Temporal extension module for skeleton-based action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 534–540. IEEE, 2021. 2, 3

[18] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021. 2

[19] Zhenyue Qin, Yang Liu, Pan Ji, Dongwoo Kim, Lei Wang, Bob McKay, Saeed Anwar, and Tom Gedeon. Fusing higher-order features in graph neural networks for skeleton-based action recognition. *arXiv preprint arXiv:2105.01563*, 2021. 2

[20] Mohamad Reza Shahabian Alashti, Mohammad Bamorovat Abadi, Patrick Holthaus, Catherine Menon, and Farshid Amirabdollahian. Rhm-har-sk: A multi-view dataset with skeleton data for ambient assisted living research. IARIA, March 2023. ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions, ACHI 2023 ; Conference date: 24-04-2023 Through 28-04-2023. 3, 4

[21] Feng Shi, Chonghan Lee, Liang Qiu, Yizhou Zhao, Tianyi Shen, Shivran Muralidhar, Tian Han, Song-Chun Zhu, and Vijaykrishnan Narayanan. Star: Sparse transformer-based action recognition. *arXiv preprint arXiv:2107.07089*, 2021. 2

[22] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3

[24] Yang Xiao, Jun Chen, Yancheng Wang, Zhiguo Cao, Joey Tianyi Zhou, and Xiang Bai. Action recognition for depth video using multi-view dynamic images. *Information Sciences*, 480:287–304, 2019. 2, 3

[25] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2

[26] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Unik: A unified framework for real-world skeleton-based action recognition. *arXiv preprint arXiv:2107.08580*, 2021. 2

[27] Dong Yang, Monica Mengqi Li, Hong Fu, Jicong Fan, and Howard Leung. Centrality graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:2003.03007*, 2020. 2

[28] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu. Learning skeletal graph neural networks for hard 3d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11436–11445, 2021. 2