Efficient Skeleton-based Human Activity Recognition in Ambient Assisted Living Scenarios with Multi-view CNN *

Mohamad Reza Shahabian Alashti[1], Mohammad Hossein Bamorovat Abadi, Patrick Holthaus, Catherine Menon and Farshid Amirabdollahian

***Abstract*— Human activity recognition (HAR) plays a critical role in diverse applications and domains, from assessments of ambient assistive living (AAL) settings and the development of smart environments to human-robot interaction (HRI) scenarios. However, using mobile robot cameras in such contexts has limitations like restricted field of view and possible noise. Therefore, employing additional fixed cameras can enhance the field of view and reduce susceptibility to noise. Nevertheless, integrating additional camera perspectives increases complexity, a concern exacerbated by the number of real-time processes that robots should perform in the AAL scenario. This paper introduces our methodology that facilitates combination of multiple views and compares different aspects of fusing information at low, medium and high levels. Their comparison is guided by parameters such as number of training parameters, floating-point operations per second (FLOPs), training time, and accuracy. Our findings uncover a paradigm shift, challenging conventional beliefs by demonstrating that simplistic CNN models outperform their more complex counterparts using this innovation. Additionally, the pivotal role of pipeline and data combination emerges as a crucial factor in achieving better accuracy levels. Ultimately, we have successfully attained a streamlined and efficient multi-view HAR pipeline, which will now be incorporated into AAL interaction scenarios.**

I. Introduction

Human activity recognition (HAR) has attract significant attention in recent years due to its wide range of applications, for example, in assistive technology or human-robot interaction (HRI) [1], [2]. The development of large datasets, multimodal sensor fusion techniques, and deep learning architectures has contributed to the progress and effectiveness of vision-based HAR systems [3], [2].

Ambient assisted living (AAL) technology has emerged as a promising approach to address the challenges faced by many older adults in maintaining their independence and quality of life [4], [5]. AAL systems are designed to be integrated into the daily environment of individuals, providing sensitive and responsive services [6]. These systems aim to aid individuals in various aspects, including handling future interfaces, recognizing frailty and mobility, preventing accidents, and supporting daily living [7], [8], [9].

In order to achieve sufficiently advanced perception capabilities, these assistive robots must be able to execute multiple tasks concurrently. Take, for example, a robot navigating an indoor environment. It must simultaneously monitor human activities, listen to commands, analyze behavior, detect and manipulate objects for the benefit of individuals, while also ensuring collision avoidance and managing unforeseen situations. This scenario exemplifies the multifaceted nature of a robot's role, involving the simultaneous execution of numerous tasks that demand efficient processing. However, this need for multitasking introduces significant computational constraints.

Furthermore, while the utilization of mobile robot cameras in these contexts offers various advantages, such as providing closer perspectives and the ability to follow humans, it also introduces certain limitations. These limitations include a restricted perspective and the possibility of noise, due to movement, in the captured data [10]. One potential solution to mitigate these challenges is the integration of additional camera perspectives alongside the robot's own viewpoint. However, when obtaining multiple perspectives from the robot, such as through cameras placed at various positions like the head and body, we encounter the drawback of heightened computational demands on an assistive robot. Potential solutions include employing high-performance machines locally within the environment or offloading computations to the cloud, exemplified by CloudMind[1] services. Another option involves utilizing static cameras within the robot workspace to capture supplementary perspectives and investigate strategies that can optimize and streamline HAR methods.

In this context, our investigation involves the evaluation of various renowned CNN models and benchmarking them with an AAL dataset, with a keen focus on essential factors, including the number of training parameters, floating-point operations per second (FLOPs), as well as training time and accuracy. Subsequently, by comparing data combination methods at low, middle and high levels, we unveil an effective and optimized multi-view architecture.

A. Research objectives

This work is focused on attaining the following objectives:

1) Improving robot perception through the development of a multi-view human activity recognition pipeline in the ambient assisted living domain.
2) Analyze the trade-offs between model complexity and accuracy by evaluating diverse CNN models on an AAL dataset and optimizing a multi-view architecture.

*Grateful for EPSRC Robot House funding (EP/P020577/1) supporting our work.

[1]All Authors are with Robotics Research Group, School of Engineering and Computer Science, University of Hertfordshire, Hatfield, United Kingdom. {m.shahabian-alashti, m.bamorovat, p.holthaus, c.menon, f.amirabdollahian2}@herts.ac.uk

[1]https://www.cloudminds.com/en

To contextualize our approach, we provide related work in Sec. II, which considers the concepts of multi-view configurations. In Sec. III, we introduce the Multi-View CNN-based HAR architectures, utilizing data from the AAL dataset. Following that in Sec. IV we examine different MV methods in terms of efficiency and effectiveness. Then, in Sec. V discuss over the results and achievements and summarise the key findings and contributions.

II. Related work

A. Multi-view Learning

The multi-view concept, as discussed in [11], involves utilizing multiple perspectives or representations of the same data to enhance the learning process. In traditional machine learning approaches, a single view of the data is used, which may limit the model's ability to capture the underlying patterns and relationships. Multi-view learning covers various aspects including representation learning [12], [13], feature selection [14], [12], and fusion methods [15], [16], [17].

The importance of the multi-view concept lies in its ability to provide a more comprehensive understanding and representation of complex and high-dimensional data [18]. By considering different views, the model can capture different aspects and nuances of the data, leading to improved accuracy [11] and generalization [19], [20]. The applications of the multi-view concept are wide-ranging. In computer vision, for example, multi-view learning can be applied to object recognition tasks, where different views of an object (e.g., different angles or lighting conditions) can be leveraged to improve recognition accuracy [12]. In natural language processing, multi-view learning can be used for sentiment analysis, where different views of textual data (e.g., word embeddings, syntactic structures) can provide a more comprehensive understanding of sentiment [13]. In this work, we augment the robotic perspective by incorporating supplementary cameras observing the same human subject engaged in indoor activities.

B. Multi-view CNNs

CNNs excel in domains like image classification [21], [22]. However, single-view CNN architectures may not fully utilize the multi-view information available in the target data. To address this limitation, multi-view architectures aim to integrate information from different views of the same data to obtain more discriminative and comprehensive representations [23].

There are two main types of multi-view CNN architectures: the one-view-one-net and the multi-view-one-net mechanisms [11]. In the one-view-one-net mechanism, each view is processed by a separate CNN, and the outputs are combined to obtain the final representation [11]. On the other hand, the multi-view-one-net mechanism models multiple feature sets together and aims to learn a common representation that captures the multi-view information [11].

Several studies have explored the application of multi-view CNNs in different domains. For example, in 3D shape recognition, multi-view CNNs have been used to model multiple views of 3D shapes and achieve better recognition performance [23]. In multivariate electroencephalography (EEG), multi-view CNNs have been employed to integrate information from multiple electrodes and improve classification accuracy [24]. Additionally, multi-view CNNs have been applied to tasks such as multi-feature aggregation [25] and lung nodule classification [26].

The design of multi-view CNN architectures involves various parameter optimization techniques [23]. For instance, the use of attention mechanisms, such as SoftPool attention, has been proposed to enhance the feature extraction and classification process [24]. Another approach is the fusion of spatial and temporal networks at different layers, which has been shown to improve performance while reducing the number of parameters [27].

Despite the strides made in multi-view CNNs, there remain critical gaps in the existing research, prompting further exploration. Firstly, there is no current comparison of various multi-view HAR approaches employing CNN methodologies to assess their performance. Furthermore, when multiple perspectives are introduced simultaneously, it substantially increases the overall complexity of the models an area. Notably, extending the concept of multi-view CNNs into the realm of activity recognition may compound this complexity. Furthermore, the research landscape has seen limited investigation into the optimal utilization of CNN models in the context of MV-HAR.

Hence, the present work seeks to bridge these gaps by presenting a comprehensive comparison of CNN methods, while investigating the effects of incorporating additional views on the model's efficiency. The aim is to formulate an optimized pipeline addressing these challenges.

III. Multi-view CNN-based HAR Architectures

In this section we systematically deconstruct the multi-view learning (MVL) structures employed and introduced in our study. Including the dataset specifications in Sec. III-A, the establishment of the multi-view CNN models Sec. III-B, feature level fusion, and the multi-view co-learning methods. Collectively, these components compose a comprehensive pipeline crucial for deploying the multi-view HAR benchmark and evaluating models across various variables.

A. AAL multi-view dataset

In order to effectively address the recognition of human activities based on skeletal data with multiple perspective in ambient assistive living scenarios, where a robot is also involved, it is crucial to select a dataset that encompasses all relevant variables. After careful consideration, we opted for the RHM-HAR-SK dataset [28], an extension of the RHM RGB data [29]. This dataset offers several advantages, primarily focusing on the classification of multi-view human activities, comprising trimmed videos from four distinct cameras: two wall-mounted (Front-view and Back-view), a mobile robot (Robot-view), and a ceiling fish-eye camera (Omni-view). These cameras' strategic placement ensures comprehensive coverage of a typical living room, creating

overlapping views. Notably, the inclusion of a robot view in this dataset renders it particularly valuable for AAL scenarios, involving situations where a robot observes and follows human activities in different locations.

A prior analysis conducted by [28] on the dataset reveals that the Omni-view exhibits low accuracy, characterized by a high number of missed poses and frames. Consequently, we opt not to incorporate the Omni-view data in our current work.

Furthermore, the RHM-HAR-SK dataset encompasses a diverse range of activity classes, a key aspect influenced by research conducted by Bedaf et al. [30]. Their study focused on identifying crucial daily activities that were essential for independent living. The dataset captures a total of fourteen daily activities captured indoor, underscoring the potential advantages that companion robots and ambient-assistive systems can offer if they can successfully detect and interpret these activities.

Capturing the spatial and temporal changes of a human skeleton within a video stream is of paramount importance, as it allows us to preserve intricate details of human body movements. We adopt a method similar to the one developed by [31], wherein the data is converted into a 3 × 34 × 34 tensor, a format carefully designed to accommodate these requirements.

B. Multi-view CNN configurations

At the heart of our framework lies the integration of information derived from diverse perspectives. To achieve this objective and gain a deeper understanding of crafting an effective and efficient structure, we adopt a systematic approach. This method involves combining data at various stages of the HAR pipeline: at the feature level Sec.III-B.1, during the batch-level training process Sec. III-B.2, at the high-level probability stage Sec. III-B.3, and through a combination of the last two levels Sec. III-B.4. In the Sec. II-B, we introduced two primary structures for Multi-View CNN, namely the "one-view-one-net" mechanism and the "multi-view-one-net" mechanism. We place specific emphasis on the "multi-view-one-net" mechanism, where all input data is learned by a single model. The lightweight and less complex nature of this approach is a crucial feature, contributing to the overall streamlined design of the framework by having only one model.

1) Multi-view low-level fusion: To implement feature level or low level fusion, the tensor data extracted from the initial stage of the HAR process is input into the classification model. In this context, our approach draws inspiration from the methodology outlined in [31], where each input camera view, represented as a tensor file as previously described in Sec. III-A, is treated as an individual input channel. In practical terms, if we consider, for instance, three camera views, the model's input configuration is adjusted to accommodate three channels, and subsequent training, validation, and testing procedures are conducted accordingly.

The advantage of this fusion method lies in its simplicity. For instance, we can apply standard image classification techniques without any modification for the three input channels, effectively processing them as RGB data. However, it's important to acknowledge the potential drawback: The fusion of input data does not encompass feature extraction beyond the initial transformation of human joints data into the tensor.

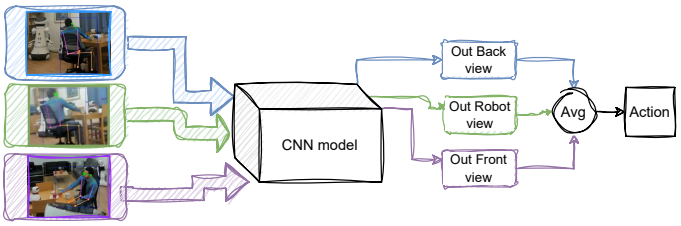

Fig. 1. The structure of high-level multi-view co-learning. Input data (34 × 34 tensor) from the same subject is sequentially fed into a single model. After obtaining the average output from all inputs (represented by a specific colour), the resultant output is used in the subsequent training or performance processes.

2) Multi-view mid-level co-learning: An approach that we introduce to combine all views involves utilizing input data from all perspectives during the training and validation process. This approach is termed*multi-view (MV) mid-level (MD) co-learning*. In other words, since we are using multi-view-one-net architecture, each batch of multi-view input data undergoes the training, validation, and testing processes collectively. Let D_R, D_B, and D_F represent the training datasets for the Robot, Back, and Front views, respectively. Each training dataset is divided into M batches. Within each batch i, and for each view j, the training process updates the model's parameters W based on the respective view's data. Since there is only one model, through each view input, the model weights will update for next view. This process repeats for all views j ($j = 1, 2, \ldots, v$) within each batch i ($i = 1, 2, \ldots, m$) for each epoch. The $\nabla L(X_j^i, W, b)$ represents the gradient of the loss function L with respect to the weights W calculated on that batch.

3) Multi-view high-level co-learning: In this MV high-level co-learning method, we combine multiple views at the highest level of the pipeline, using only the average output in the learning process. This is distinct from mid-level co-learning, which involves individual view outputs in the learning process. However, similar to mid-level co-learning, this method follows the multi-view-one-net structure, as depicted in Figure 1 for MV-HG co-learning.

Let $V_1, V_2, \ldots, V_v$ represent the views (e.g., Robot, Back, Front) for a given dataset. Each view has a set of outputs, $O_1, O_2, \ldots, O_v$, where O_j is the set of outputs for view V_j. For the MV High-level Co-learning approach, the learning process involves averaging the predictions from all views. This can be mathematically expressed as:

$$O_{\text{avg}} = \frac{1}{v} \sum_{j=1}^{v} O_j \tag{1}$$

Where:

- O_{avg} represents the average predictions for all views.

- v is the total number of views.
- O_j is the set of predictions for view V_j.

This process involves obtaining the average predictions, O_{avg}, which is then used for further learning and model updates.

4) Combining multi-view mid-level and high-level co-learning: The limitation in the employment of the MD method manifests in the inability to concurrently harness the outputs from all perspectives within the one-net structure. However, leveraging the multi-view-one-net architecture in both mid-level and high-level co-learning methodologies enables us to combine the advantages of both approaches. We first employ mid-level training to enhance the model's training and validation process and subsequently perform test set classification using the high-level combination. This strategy involves training and validating the model with a more diverse data and testing it using the average of predictions. This approach enhances the single model's training process by enriching it with diverse input data through mid-level co-learning. It further capitalizes on high-level co-learning during the model's performance stage. To prevent the over-fitting issue the test data has been separated from the training and validation in both Mid and high level co-learning methods.

IV. Experiments

This section presents a number of experiments and their results obtained as well as analysis focusing on comparative model performance. As established state-of-the art and each representing a distinct architecture, we compare the following CNN models in our experiment: LeNet, M-LeNet, ResNet18, MobileNet, SqueezNet, DenseNet, and MnasNet. To measure model performance as a dependent variable, we compared the recognition *accuracy*, model *complexity* (number of parameters), *computational complexity* (FLOPs), and *performance time* as individual metrics. To assess potential influences on MV frameworks' performance, we considere the Robot view as the base single view for improvement, along with low-level fusion (LW), mid-level co-learning (MD), high-level co-learning (HG), and the combined MD and HG (MH) methods as additional independent variables in comparing MV methods.

A. Experimental Settings

We performed individual tests for CNN models under identical conditions, randomly sampling 34 frames from the dataset RHM-HAR-SK [28], which integrates Yolo7 pose estimation to extract human skeleton data. Hyperparameters encompass a dropout rate of 0.0135, we set the learning rate to 0.000072, with a weight decay of 0.0000521, a batch size of 128, and a training duration spanning 128 epochs. We employed the Adam [32] optimization method, and the experimental design incorporates stratified K-Fold cross-validation with five folds to ensure a uniform distribution of classes in each fold. Throughout the training, model performance is assessed on a validation set after each epoch, utilizing the cross entropy loss as the chosen loss function. The dataset is split into training and validation sets during cross-validation, with 20% of the data reserved for subsequent testing. We evaluate the trained model on the test set, and key metrics, including test loss, accuracy, precision, recall, F1-score, and the confusion matrix, are logged for analysis. We performed tests on a high-performance computer with the following specifications:

- Architecture: $X86\_64$
- CPU: AMD Ryzen Threadripper PRO 5975WX (32 Cores, 64 Threads, 7006.64 MHz)
- Graphics Card: NVIDIA GA102GL [$RTXA6000$] (10,752 CUDA cores, memory size of 48 GB GDDR6)
- Storage: PC801 NVMe SK hynix 2TB SSD
- Memory: 125 GiB RAM

B. Model complexity & HAR

In this segment of our experiment, we explore the intricate relationship between model complexity and human activity recognition (HAR). This experiment challenges the conventional belief that more complex CNN models inherently yield superior accuracy. To test this hypothesis, we rigorously train and test various CNN models, focusing initially on single-view HAR with the robot view as the base perspective. The results, as depicted in Table IV-B shows that models with fewer parameters and complexity, such as MnasNet (77.72%) and M-LeNet (77.14%), not only hold their own but also outperform single-view models even with higher parameters and complexity, like ResNet (76.91%) and DenseNet (74.81%).

C. Views & HAR accuracy

When analysing the impact of additional views on HAR accuracy, our primary goal is to see the effect of enhancing a robot's perception by introducing multiple external cameras. The results in Sec. IV-B reveal that, except for some models in the LW method, all other view combination frameworks have improved the accuracy, showcasing the potential of multi-view perspectives.

The MH method notably demonstrates the highest improvement, ranging from 10% to 25% across various models. The top four highest accuracies are consistently achieved using the MH method, with notable performances from ResNet (90.90%), MnasNet (90.08%), DenseNet (89.93%), and M-LeNet (88.59%).

Conversely, the HG method showcases improvement of 5 to 14 percent, with ResNet (85.25%), M-LeNet (83.95%), and DenseNet (83.94%) securing the top spots in terms of accuracy. Notably, SqueezeNet exhibits a remarkable jump in accuracy within the HG framework.

For the MD method, improvements are slightly lower, ranging from 1 to 10.5 percent, with MnasNet (82.28%), DenseNet (81.05%), and M-LeNet (80.14%) claiming the top accuracy spots in the MD method. However, the LW framework's results are less consistent, with some models experiencing improvement and others exhibiting declines. Despite challenges, certain models like MobileNet (76.91%) and ResNet (76.91%) demonstrate reasonable improvement within the LW framework.

Models	Acc. (%)	#Params (M)	FLOPs(G)
LeNet-LW	75.91	0.062006	0.00075
LeNet-HG	79.57	0.061706	0.00048
LeNet-MD	75.87		
LeNet-MH	84.71	-	-
LeNet-R	74.74		
M-LeNet-LW	76.67	0.621364	0.00125
M-LeNet-HG	**83.95**	**0.621184**	**0.00106**
m-LeNet-MD	80.14		
M-LeNet-MH	**88.59**	-	-
M-LeNet-R	**77.14**		
ResNet18-LW	82.04	11.689512	0.08102
ResNet18-HG	**85.25**	11.177422	0.07870
ResNet18-MD	77.52		
ResNet18-MH	**90.90**	-	-
ResNet18-R	76.91		
MobileNet-LW	76.91	3.504872	0.01644
MobileNet-HG	70.53	2.24123	0.01501
MobileNet-MD	65.18		
MobileNet-MH	**83.59**	-	-
MobileNet-R	58.76		
SqueezNet-LW	55.47	1.235496	0.00791
SqueezNet-HG	76.38	0.728526	0.00559
SqueezNet-MD	73.25		
SqueezNet-MH	**86.73**	-	-
SqueezNet-R	62.68		
DenseNet-LW	70.72	7.978856	0.06677
DenseNet-HG	**83.94**	6.961934	0.06395
DenseNet-MD	81.05		
DenseNet-MH	**89.93**	-	-
DenseNet-R	74.81		
MnasNet-LW	35.32	2.218512	0.00702
MnasNet-HG	81.51	**2.138512**	**0.00622**
MnasNet-MD	82.28		
MnasNet-MH	**90.08**	-	-
MnasNet-R	**77.72**		

TABLE I

SUMMARY OF PERFORMANCE METRICS FOR ANALYSED CNN MODELS IN HUMAN ACTIVITY RECOGNITION, LISTING ACCURACY AND MODEL COMPLEXITY EXPRESSED IN NUMBER OF PARAMETERS AND FLOATING-POINT OPERATIONS. SUFFIX STANDS FOR, LW: LOW-LEVEL FUSION MD: MID-LEVEL CO-LEARNING HG: HIGH-LEVEL CO-LEARNING MH: COMBINED MID-LEVEL AND HIGH-LEVEL (MD AND HG) CO-LEARNING

D. CNNs in multi-view trade-offs

In this phase of the experiment, we investigate the variation in the trade-off between model accuracy and complexity across different CNN architectures in the context of multi-view HAR. Specific CNN architectures may demonstrate superior trade-offs between accuracy and model complexity in the multi-view HAR scenario. To identify the optimal framework, we assess the accuracy density [33], [34], calculated as the ratio of accuracy to the number of parameters. A higher accuracy density signifies greater efficiency. In Figure 2 we plot the top accuracy of each tested model (MH method) against their accuracy density. The accuracy density serves as a metric to gauge the efficiency of parameter utilization for each model. In this figure, the run time of each model during the training and validation processes is depicted using

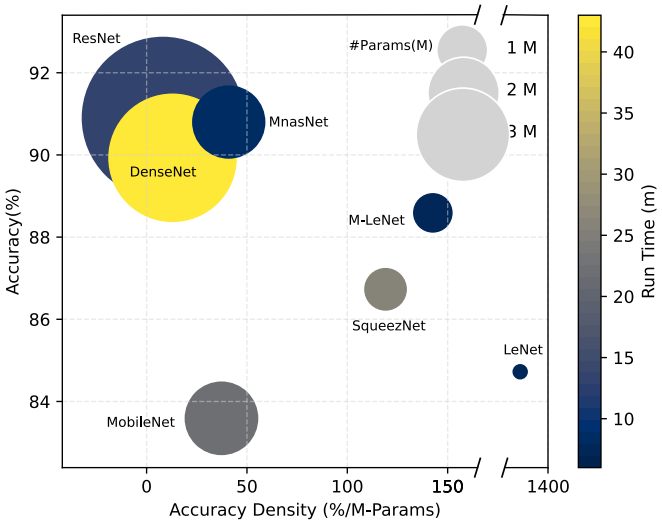

Fig. 2. Accuracy density analysis for CNN models in the MH method. The centre of the circles represents the relevant value on the axes, denoting accuracy and accuracy density. Circle diameters correlate with the number of parameters, while their colours reflect the time spent during the training process.

colour mapping, varying between 6 minutes (blue) for LeNet and 43 minutes for DenseNet (yellow).

The findings indicate that simpler models like LeNet exhibit greater effectiveness and efficiency compared to more complex CNN architectures such as MobileNet in the HAR task. Despite LeNet achieving an accuracy density of nearly 1400, making it a more effective model, its accuracy falls approximately 5% lower than the highest-performing model (ResNet at 90.9%). However, the top three accuracy models (ResNet, MnasNet, and DenseNet) display lower efficiency, with an accuracy density lower than 50. Notably, among these, DenseNet emerges as the slowest model in terms of performance, requiring approximately 40 minutes for training. Additionally, while M-LeNet and SqueezeNet demonstrate similar levels of efficiency, the runtime of SqueezeNet is nearly three times longer than that of M-LeNet. This observation highlights the direct impact of FLOPs on model performance time, considering that the FLOPs value in M-LeNet is approximately five times smaller than that in SqueezeNet.

V. DISCUSSION AND CONCLUSIONS

This work presented a multi-view CNN-based human activity recognition architecture and showed that additional camera views can efficiently enhance robot perception in ambient assisted living scenarios. The trade-offs between model complexity and accuracy reveals that lightweight CNN models like M-LeNet and MnasNet are more efficient than complex architectures, challenging conventional beliefs.

We selected the RHM-HAR-SK dataset capturing diverse human activities from four camera angles. Converting human skeleton stream data into a tensor format allowed structured representation of spatial and temporal changes. The utilization of this input format empowered us to combine informa-

tion seamlessly through fusion, co-learning, and combination mechanisms within the "multi-view-one-net" architecture.

In our proposed architectures, notably the MH method, the integration of mid-level and high-level co-learning, exhibits promising outcomes across various CNN models, underscoring the considerable potential of incorporating multi-view information for HAR tasks. This combination surpasses the constraints associated with mid-level methodologies, where obtaining individual views simultaneously poses challenges. It leverages the generalization capabilities of mid-level methods in the training phase, while simultaneously harnessing the power of the high-level method through the aggregation of average outputs during performance evaluation.

Our results demonstrate that a locally efficient multi-view HAR model in a robot can effectively operate in ambient assisted living scenarios. This effectiveness is achieved by selecting an appropriate model like M-LeNet with low complexity and a limited number of parameters, coupled with a well-designed multi-view structure.

Methods for 2D pose extraction that can handle concurrent processing of multiple individuals without overloading are scalable, enabling efficient execution of multiple HAR tasks. In future work, we will look into streamlining the skeleton-based model for multi-person HAR deployment, ensuring scalability to manage multiple tasks efficiently.

REFERENCES

[1] S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," *International Journal of Distributed Sensor Networks*, vol. 12, no. 8, p. 1550147716665520, 2016.

[2] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106970, 2021.

[3] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[4] F. Amirabdollahian, R. o. d. Akker, S. Bedaf, R. Bormann, H. Draper, V. Evers, J. G. Pérez, G. J. Gelderblom, C. G. Ruiz, D. Hewson *et al.*, "Assistive technology design and development for acceptable robotics companions for ageing years," *Paladyn, Journal of Behavioral Robotics*, vol. 4, no. 2, pp. 94–112, 2013.

[5] S. Blackman, C. Matlo, C. Bobrovitskiy, A. Waldoch, M. L. Fang, P. Jackson, A. Mihailidis, L. Nygård, A. Astell, and A. Sixsmith, "Ambient assisted living technologies for aging well: a scoping review," *Journal of Intelligent Systems*, vol. 25, no. 1, pp. 55–69, 2016.

[6] T. Kleinberger, M. Becker, E. Ras, A. Holzinger, and P. Müller, "Ambient intelligence in assisted living: enable elderly people to handle future interfaces," in *Universal Access in Human-Computer Interaction. Ambient Interaction: 4th International Conference on Universal Access in Human-Computer Interaction, UAHCI 2007 Held as Part of HCI International 2007 Beijing, China, July 22-27, 2007 Proceedings, Part II 4*. Springer, 2007, pp. 103–112.

[7] M. Trombini, F. Ferraro, M. Morando, G. Regesta, and S. Dellepiane, "A solution for the remote care of frail elderly individuals via exergames," *Sensors*, vol. 21, no. 8, p. 2719, 2021.

[8] K. Denecke, "What characterizes safety of ambient assisted living technologies?" *European Federation for Medical Informatics (EFMI) and IOS Press*, 2021.

[9] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE journal of biomedical and health informatics*, vol. 17, no. 3, pp. 579–590, 2012.

[10] Y. Jang, I. Jeong, M. Younesi Heravi, S. Sarkar, H. Shin, and Y. Ahn, "Multi-camera-based human activity recognition for human–robot collaboration in construction," *Sensors*, vol. 23, no. 15, p. 6997, 2023.

[11] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, 2021.

[12] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.

[13] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *Ieee Access*, vol. 7, pp. 63 373–63 394, 2019.

[14] C. Ahuja, L. P. Morency *et al.*, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, pp. 1–20, 2017.

[15] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[16] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[17] S. Sun, "A survey of multi-view machine learning," *Neural computing and applications*, vol. 23, pp. 2031–2038, 2013.

[18] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," *Information Fusion*, vol. 50, pp. 158–167, 2019.

[19] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," *arXiv preprint arXiv:1702.02519*, 2017.

[20] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," *arXiv preprint arXiv:2002.07017*, 2020.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[22] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view cnns for object classification on 3d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5648–5656.

[23] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 945–953.

[24] W. Wang, X. Wang, G. Chen, and H. Zhou, "Multi-view softpool attention convolutional networks for 3d model classification," *Frontiers in Neurorobotics*, vol. 16, p. 1029968, 2022.

[25] W. Zou, D. Zhang, and D.-J. Lee, "A new multi-feature fusion based convolutional neural network for facial expression recognition," *Applied Intelligence*, vol. 52, no. 3, pp. 2918–2929, 2022.

[26] K. Liu and G. Kang, "Multiview convolutional neural networks for lung nodule classification," *International Journal of Imaging Systems and Technology*, vol. 27, no. 1, pp. 12–22, 2017.

[27] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[28] M. R. S. Alashti, M. B. Abadi, P. Holthaus, C. Menon, and F. Amirabdollahian, "Rhm-har-sk: A multi-view dataset with skeleton data for ambient assisted living research," *IARIA, March*, 2023.

[29] M. Bamorovat Abadi, M. R. Shahabian Alashti, P. Holthaus, C. Menon, and F. Amirabdollahian, "Rhm: Robot house multi-view human activity recognition dataset," in *ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions*. IARIA, 2023.

[30] S. Bedaf, G. J. Gelderblom, D. S. Syrdal, H. Lehmann, H. Michel, D. Hewson, F. Amirabdollahian, K. Dautenhahn, and L. De Witte, "Which activities threaten independent living of elderly when becoming problematic: inspiration for meaningful service robot functionality," *Disability and Rehabilitation: Assistive Technology*, vol. 9, no. 6, pp. 445–452, 2014.

[31] M. R. S. Alashti, P. Holthaus, C. Menon, and F. Amirabdollahian, "Lightweight human activity recognition for ambient assisted living,"," *IARIA, March*, 2023.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.

[34] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE access*, vol. 6, pp. 64 270–64 277, 2018.