Robotic Vision and Multi-View Synergy: Action and activity recognition in assisted living scenarios

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

Abstract— **In the realm of integrating robotics within human-centric settings, the significance of Human-Robot Interaction (HRI) is increasingly evident. A crucial component of effective HRI is Human Activity Recognition (HAR), which is instrumental in enabling robots to respond aptly in human presence, especially within Ambient Assisted Living (AAL) environments. Since robots are generally mobile and their visual perception is often compromised by motion and noise, this research evaluates methods by merging the robot's mobile perspective with a static viewpoint in multi-view deep learning models. We introduce a dual-stream Convolutional 3D (C3D) model aimed at improving vision-based HAR accuracy for robotic applications. Utilising the Robot House Multiview (RHM) dataset, which encompasses a robotic perspective along with three static views (Front, Back, Top), we examine the efficacy of our model and conduct comparisons with the dual-stream ConvNet and SlowFast models. The primary objective of this study is to enhance the accuracy of robot viewpoints by integrating them with static views using dual-stream models. The metrics for evaluation include Top-1 and Top-5 accuracy. Our findings reveal that the integration of static views with robotic perspectives significantly boosts HAR accuracy in both Top-1 and Top-5 metrics across all models tested. Moreover, the proposed dual-stream C3D model demonstrates superior performance relative to other contemporary models in our evaluations.**

I. Introduction

Human-robot interaction (HRI) is becoming increasingly vital in Ambient Assisted Living (AAL). This trend emphasises the importance of robots to blend smoothly into human environments. Integration of robots in these settings goes beyond basic task performance. It involves complex social and emotional interactions, considerably expanding the HRI field [1], [2]. Understanding human behaviours deeply is critical in advancing HRI. Therefore, Human Action Recognition (HAR) is a key component in this area. Recent advancements in Machine Learning and Deep Learning have significantly improved HAR's efficiency. These developments are crucial for creating intelligent robots. Such robots need to be responsive and adaptable in human-centred environments [3], [4]. These technological advancements are poised to transform assistive robotics. They offer the potential to improve life quality and autonomy for those in need of assistance [5].

Despite recent progress, achieving precise Human Action Recognition (HAR) in the dynamic and unpredictable environments where robots operate remains a considerable challenge. A key issue is the limitations in robot view datasets. These datasets, although dynamic, do not match the quality of static, human-centric views in terms of scope and accuracy [6], [7]. In 2022 and 2023, research studies such as those conducted by Smith et al. [8] and Jones and Kumar [9] highlighted these difficulties. These studies introduced new methods aimed at improving the accuracy and reliability of HAR within robotic systems.

To tackle these challenges, our research introduces an innovative method that integrates robot views with static views from the Robot House Multiview (RHM) dataset [6], [10]. We use a dual-stream network that incorporates the Convolutional 3D (C3D) model. This approach effectively combines these diverse perspectives. The technique aims to significantly enhance the robot's perception accuracy. It provides a more thorough understanding of the surrounding environment. Our method specifically focuses on analyzing spatial frames in both streams. This analysis is essential for capturing the essence of both dynamic and static elements present in the robot's operational environment.

Additionally, our research extends to assess the effectiveness of other well-known models like the dual-stream ConvNet [11] and Slowfast [12] in similar settings. Recent research, including works by Garcia and Lee [13] and Zhang et al. [14], has demonstrated the capabilities of these models in processing complex visual data. This makes them suitable for comparison in our study. We anticipate that incorporating these varied approaches will provide substantial insights. These insights are expected to contribute to enhancing the performance of robots in Human Action Recognition (HAR) tasks. Also, the significance of incorporating temporal information is thoroughly discussed in [15].

The structure of this paper is organised as follows: Section II provides a comprehensive review of the existing literature on multi-stream models, establishing a crucial background for our proposed methodology. Section III elaborates on our experimental approach, detailing the technical implementation of the dual-stream C3D model. In Section IV, we present and analyze the results of our experiments, emphasising the improvements achieved through our approach. Finally, Section V concludes the paper. This section summarises our main findings and discusses their potential impact on future research and practical applications in Human Action Recognition (HAR) within Ambient Assisted Living (AAL) scenarios.

II. Related Work

Enhancing Human Action Recognition (HAR) can be achieved through the use of multi-stream networks. These

[1] All Authors are with the School of Physics, Engineering and Computer Science., University of Hertfordshire, UK m.bamorovat@herts.ac.uk

deep learning models are designed to recognise human actions by processing various types of data simultaneously. This data includes skeletal formations, motion information, and object interactions [16]. Multi-stream network models often employ techniques like Motion History Images (MHI) or Optical Flow to capture and analyse temporal information. By adopting this comprehensive approach, these networks can understand human actions through a detailed consideration of contextual, global, and local motion attributes [17]. This section provides an overview of multi-stream networks, which form the foundational approach of our research in integrating two views into a cohesive dual-stream model.

In a foundational study on multi-view networks for Human Action Recognition (HAR), Wang et al. [18] developed an approach using deep dual-stream Convolutional Networks (ConvNets). They complemented this with Fisher vector-encoded Improved Dense Trajectories (IDT) features [19]. In their methodology, temporal features were processed in a distinct stream. This approach yielded highly effective results, marking a significant advancement in the field of HAR.

Karpathy et al. [20] made significant contributions to the fusion aspect of multi-stream networks. Their research delineates that in multi-stream network architectures, the integration of results from different streams can be categorised into four main approaches: early fusion, mid-level fusion, late fusion, and lateral fusion. In early fusion, features from multiple streams are combined at an initial stage before they are inputted into the network for final classification. Mid-level fusion involves merging intermediate features from each stream at a midpoint in the process, before making the final prediction. Late fusion, on the other hand, employs a weighted average of predictions from each stream to arrive at the final result. Lastly, lateral fusion operates by processing the streams in parallel while intermittently sharing features or information between them.

In a subsequent study by Wang et al. [21], they further advanced Human Action Recognition (HAR) using 'very deep dual-stream ConvNets'. This approach was inspired by prominent models such as GoogLeNet and VGGNet. Their focus was on dual-stream processing, specifically for video frames and motion data. They implemented various strategies aimed at preventing overfitting, particularly in smaller datasets. Additionally, they enhanced the Caffe toolbox to boost overall performance. Their findings demonstrated that dual-stream models are particularly effective, even in scenarios involving smaller datasets. This insight was a significant contribution to the field of HAR.

Feichtenhofer et al. [22] conducted a detailed exploration of various fusion methods to enhance Human Action Recognition (HAR) in videos using Convolutional Networks. They discovered that fusing features at the convolutional layer, rather than at the softmax layer, maintained high-performance levels while utilising fewer parameters. Furthermore, they noted that combining features at both the final convolutional layer and the class prediction layer significantly improved accuracy. Their experiments also included temporal fusion methods such as 3D pooling and 3D filtering, which resulted in better recognition accuracy. By integrating this Convolutional-based approach with Improved Dense Trajectories (IDT) handcrafted features through late fusion, their method exceeded the performance of both the original dual-stream model and other existing techniques, marking a notable advancement in the field.

Wang et al. [23] addressed a key limitation of traditional dual-stream Convolutional Networks (ConvNets), which struggled to capture complex spatial and temporal information effectively. To overcome this challenge, they developed the Spatiotemporal Pyramid Network, a novel approach in the field. This network incorporates a unique Spatiotemporal Compact Bilinear (STCB) operator. The STCB operator is designed to fuse spatial and temporal features more efficiently. The results from their study demonstrated that the Spatiotemporal Pyramid Network surpasses the performance of previous methods in Human Action Recognition (HAR), showcasing its effectiveness in handling intricate spatiotemporal data.

Feichtenhofer et al. [24] were pioneers in integrating deep Residual Networks (ResNets) into multi-stream networks. They introduced an innovative Convolutional Networks (ConvNets) architecture specifically tailored for video-based Human Action Recognition (HAR). This architecture marked a significant departure from traditional dual-stream designs, focusing instead on multiplicative spacetime feature interactions. Central to their model is the use of deep ResNets, which are further enhanced through the incorporation of cross-stream residual connections. This novel approach represented a key advancement in the field, offering a more sophisticated method for handling the complex demands of video-based HAR.

Zhu et al. [25] introduced a groundbreaking approach to decrease the computational burden associated with extracting temporal features in Human Action Recognition (HAR). They developed a novel Convolutional Neural Network (CNN) architecture known as "hidden dual-stream networks." This innovative architecture is designed to capture motion between video frames directly. A key advantage of this method is that it eliminates the need for pre-computed optical flow, thereby allowing for end-to-end training of the network. The results of their study were significant, demonstrating that this approach could outperform leading methods in real-time action recognition. This advancement not only enhances efficiency but also opens up new possibilities for real-time HAR applications.

Building on the findings of Karpathy [20] and addressing the challenge of removing handcrafted temporal features, Feichtenhofer et al. [12] made a significant contribution with their development of the SlowFast networks for video recognition. This innovative architecture consists of two distinct pathways: a 'Slow' pathway designed to capture spatial details and a 'Fast' pathway focused on temporal motion. A key feature of this architecture is the incorporation of lateral connections between these pathways, which facilitates a more effective integration of spatial and temporal information. The performance of the SlowFast networks was noteworthy, as they demonstrated superior capabilities, achieving substantial accuracy improvements in video-based Human Action Recognition (HAR). This approach marked a significant advancement in the field, enhancing the efficiency and effectiveness of HAR systems.

We hypothesise that the use of dual-stream networks, augmented with the integration of static views from multiview datasets such as RHM [6], can enhance the accuracy of robot's perspective in Human Action Recognition (HAR). The Robot House Multi-View (RHM) dataset encompasses four distinct perspectives: Front, Back, Ceiling (Omni), and robot-views, incorporating a comprehensive collection of 14 classes across 6701 video clips for each view, culminating in a total of 26,804 video clips across all views. Each video clip spans a duration ranging from 1 to 5 seconds. Notably, video clips bearing identical numerical identifiers and class allocations are synchronized across the different views, ensuring consistency in multi-view analysis.

To test this, we have built upon the dual-stream CNN architecture established by Simonyan et al. [11], and the lateral fusion techniques described by Karpathy et al. [20] and Feichtenhofer et al. [12]. Central to our approach is the incorporation of the foundational 3D CNN (C3D) model by Tran et al. [26]. This integration has culminated in the creation of a novel dual-stream model, which is elaborated in Section III of our research. This model represents a significant step forward in the application of dual-stream networks for improving robotic perception in HAR scenarios.

TABLE I

Dual-Stream C3D Model Description, 'I' represents the Input to each layer. 'O' denotes the Output, indicating the number of filters in the layer. 'K' refers to the Kernel size. 'P' illustrates the Padding size, and 'S' signifies the Stride size of each layer.

Stage	First Stream	Second Stream
Data Layer	16*112*112	16*112*112

Conv$_1$	I=3, O=32
	K=(3,3,3), P=(1,1,1)
		I=3, O=32
		k=(3,3,3), p=(1,1,1)

pool$_1$	K=(1,2,2), S=(1,2,2)	K=(1,2,2), S=(1,2,2)

Conv$_2$	I=32, O=64
	K=(3,3,3), P=(1,1,1)
		I=64, O=64
		K=(3,3,3), P=(1,1,1)

pool$_2$	K=(2,2,2), S=(2,2,2)	K=(2,2,2), S=(2,2,2)

Conv$_{3a}$	I=64, O=128
	K=(3,3,3), P=(1,1,1)
		I=128, O=128
		K=(3,3,3), P=(1,1,1)

Conv$_{3b}$	I=128, O=128
	K=(3,3,3), P=(1,1,1)
		I=128, O=128
		K=(3,3,3), P=(1,1,1)

pool$_3$	K=(2,2,2), S=(2,2,2)	K=(2,2,2), S=(2,2,2)

Conv$_{4a}$	I=128, O=256
	K=(3,3,3), P=(1,1,1)
		I=256, O=256
		K=(3,3,3), P=(1,1,1)

Conv$_{4b}$	I=256, O=256
	K=(3,3,3), P=(1,1,1)
		I=256, O=256
		K=(3,3,3), P=(1,1,1)

pool$_4$	K=(2,2,2), S=(2,2,2)	K=(2,2,2), S=(2,2,2)

Conv$_{5a}$	I=256, O=256
	K=(3,3,3), P=(1,1,1)
		I=512, O=512
		K=(3,3,3), P=(1,1,1)

Conv$_{5b}$	I=256, O=256
	K=(3,3,3), P=(1,1,1)
		I=512, O=512
		K=(3,3,3), P=(1,1,1)

pool$_5$	K=(2,2,2), S=(2,2,2)	K=(2,2,2), S=(2,2,2)
Concatenate & FC6 & FC7		Classes

III. Methodology

In our research, we have innovated a new dual-stream model by integrating key architectural concepts from the domain of Human Action Recognition. The foundational structure of our model is based on the Convolutional 3D (C3D) model. We use the C3D model to generate both streams in our dual-stream framework. A crucial aspect of our model is the incorporation of lateral connections between these streams. These lateral connections are instrumental in facilitating an efficient exchange of information between the streams, thereby enriching each stream with insights gained from the other. For the fusion process within the lateral connections, we employ concatenation. This method of integration has proven effective in enhancing the overall performance of the model. The total number of parameters in this newly developed dual-stream model is approximately 92.81 million, reflecting the complexity and robustness of the architecture.

The detailed architecture of our dual-stream C3D model is depicted in Figure 1. This illustration provides a visual representation of the model's structure, showcasing the integration of the two streams and the implementation of lateral connections. Additionally, specific information regarding the model is presented in Table I. This table includes comprehensive details about the model, such as the number of parameters and the configurations of various layers.

The details of the Figure 1 are as follow:

- The network is composed of two streams: the upper stream (first stream) and the lower stream (second stream).
- Convolution layers are represented by yellow boxes within each stream.
- Padding layers are indicated by orange boxes.
- The network includes two fully connected layers and softmax layers for classification tasks.
- Lateral connections between the streams feature concatenation fusion for integrating information.

IV. Experiments and results

To investigate the effect of integrating multi-view data with a robot perspective on the dual-stream C3D model, a series of experiments was conducted. The primary goal of these experiments was to enhance the performance of the robot view. Consequently, all experimental designs were centred around this perspective.

Initially, the proposed model underwent evaluation using a combination of different views alongside the robot view. This step was crucial to assess how various perspectives complement the robot view within the model. In addition, to gain a deeper understanding of the impact of multi-view data on dual-stream models, we conducted tests that focused on spatial input data. This approach was chosen to specifically exclude the influence of temporal information. Therefore, in

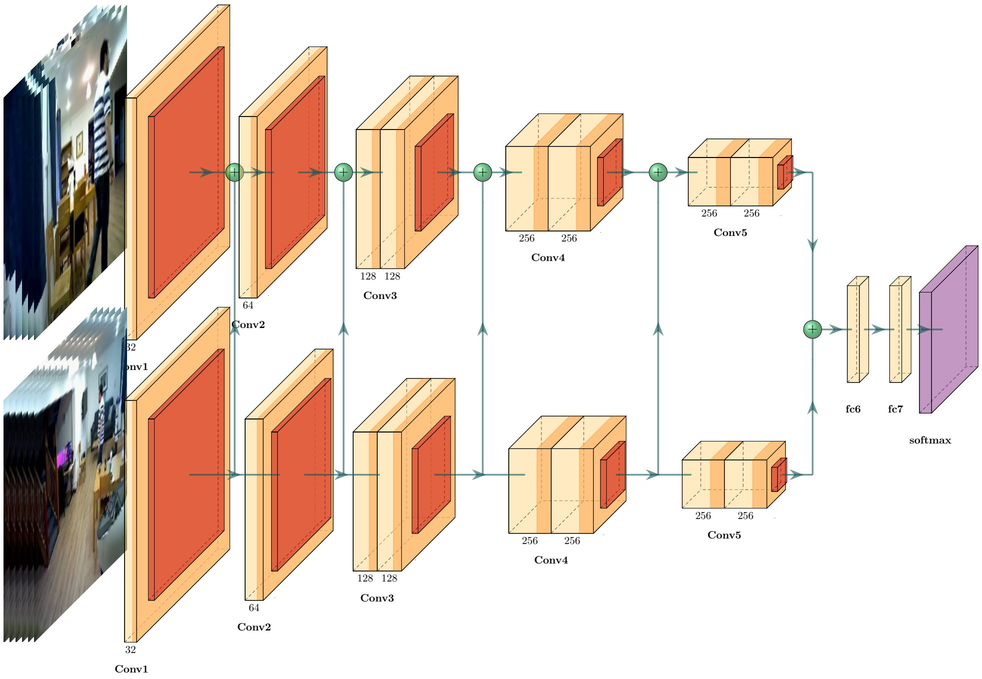

Fig. 1. **Dual-Stream C3D Network Architecture**, This design includes an upper (first) and a lower (second) stream. Key features are yellow-boxed convolution layers and orange-boxed padding layers. The architecture integrates two fully connected softmax layers for classification. Lateral connections with concatenation fusion link the streams.

this phase of testing, only spatial frames were used in both streams of the model.

To thoroughly compare the impact of multi-view data on dual-stream networks and to rigorously evaluate our proposed model, we conducted a series of identical experiments with well-established models: the dual-stream ConvNet [11] and the SlowFast model [12].

For these comparative experiments, we used the Robot House Multiview (RHM) dataset, ensuring consistency across all tests. The training parameters were standardised for all models to ensure a fair comparison. We set a batch size of 30 and a frame count of 16 for processing. The learning rate was fixed at 0.0001, and we employed the Stochastic Gradient Descent (SGD) optimiser for training.

In terms of performance evaluation, we chose top-1 and top-5 accuracy as our primary metrics. These metrics are widely recognised in the field and provide a clear measure of the model's ability to accurately classify actions from the provided data. By using these standardised settings and evaluation criteria, we aimed to achieve a comprehensive and objective comparison of the models' performances in handling multi-view data, particularly focusing on the enhancement offered by our proposed dual-stream C3D model.

A. *Single views input*

In our initial experiments, we tested the dual-stream model using identical frames in both streams to understand its effect on each view and establish a reference for the impact of multi-view data. We employed specific viewpoint pairs including robot-robot, front-front, back-back, and top-top configurations, allowing us to analyse the model's performance under viewpoint uniformity. These tests were crucial for assessing the baseline capabilities of the dual-stream model before introducing multi-view variations. The results of these experiments, which provide insights into the model's performance with uniform viewpoints, are compiled in Table II. This data serves as a foundational reference for further multi-view impact analysis on the dual-stream model.

The experimental results demonstrate that our dual-stream C3D model significantly outperforms the original single-stream C3D model, as detailed in [6], in both Top-1 and Top-5 accuracy metrics across various viewpoints. Notably, for the Robot view, the dual-stream C3D model exhibits a remarkable 10% increase in Top-1 accuracy. Similarly, there is a 1% increase in Top-1 accuracy for the Front view, a 1% rise for the Back view, and a 2% improvement for the Top view.

Furthermore, the proposed dual-stream C3D model shows superior performance in all viewpoints compared to other models, as outlined in Table II. It surpasses the SlowFast model by a significant margin of more than 15% and exhibits an improvement of over 5% compared to the dual-stream ConvNet model across all views. Interestingly, the Front view consistently yielded the highest results, while the Robot view recorded the lowest performance in all three models, providing valuable insights into viewpoint-specific model efficacy.

TABLE II

Performance of dual-stream Models for Identical Viewpoints in the RHM Dataset. In this setup, both streams of the dual-stream model (* DS: Dual-Stream) process the same frame from the same viewpoint. The input frames are spatial, representing normal visual data.

Inputs		SlowFast (101)		DS* ConvNet		DS* C3D	
Second stream	First stream	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Robot View	Robot View	42.24	88.19	48.26	89.02	**54.91**	**89.16**
Front View	Front View	58.63	95.43	63.82	96.38	**68.05**	**98.26**
Back View	Back View	57.87	95.68	62.59	96.27	**67.13**	**98.17**
Top View	Top View	54.39	95.39	60.44	95.98	**64.80**	**97.17**

TABLE III

Results of the dual-stream Model with Multi-View Input. In this configuration, each stream of the dual-stream model (* DS: Dual-Stream) processes the same frame number, but from different camera viewpoints. The frames used are normal, indicating that they consist of spatial data. The table compares the performance of three different models under these conditions.

Inputs		SlowFast (101)		DS* ConvNet		DS* C3D	
Second stream	First stream	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Front View	Robot View	45.28	91.31	62.77	94.51	**71.06**	**98.14**
Back View	Robot View	44.69	90.64	61.02	93.89	**66.25**	**97.17**
Top View	Robot View	44.91	87.75	59.76	92.21	**67.91**	**97.2**
Robot View	Front View	41.86	89.95	58.77	91.98	**65.09**	**95.95**
Robot View	Back View	40.87	88.59	57.51	91.70	**62.7**	**94.42**
Robot View	Top View	40.27	88.02	56.68	90.79	**64.6**	**95.7**

B. *Multi-views input*

Intending to assess the impact of multi-view data on dual-stream models, especially in terms of enhancing robot view performance, we methodically designed our experiments to incorporate the robot view.

In our first experimental series, the dual-stream model was configured with the robot view as the primary stream. This configuration tested combinations such as robot-front, robot-back, and robot-top pairings. The intent was to understand how the robot view interacts and contributes to the overall performance when it is the leading perspective in the model.

Subsequently, in our second series of experiments, we shifted the robot view to the secondary stream. This change in setup included pairing combinations like front-robot, back-robot, and top-robot. This arrangement allowed us to evaluate the model's performance when the robot view complements the primary view.

The results from both experimental series are comprehensively compiled in Table III. This table provides a detailed overview of the performance of each viewpoint pairing. It specifically highlights how the integration of the robot view in various configurations influences the accuracy of the models, including the SlowFast, dual-stream ConvNet, and our dual-stream C3D models. These results are instrumental in understanding the effectiveness of multi-view data in enhancing the performance of dual-stream models, particularly from a robot-centric perspective.

The results from our experiments demonstrate that the inclusion of additional viewpoints as separate streams in the dual-stream model significantly enhances the accuracy of the Robot View. In all six viewpoint combinations, there was a notable improvement in the performance of the Robot View within the dual-stream framework. Particularly striking was the increase in accuracy when the Robot View served as the primary stream, fused with other views. This finding suggests that the additional information provided by the multi-view setup strengthens the model's capacity to distinguish between different actions.

However, when the Robot View was configured as the secondary stream, there was a relative decrease in accuracy compared to the setups where it was the primary input. This observation implies that the contribution of the Robot View's information as a secondary input has a lesser impact on the overall model performance.

In every pairwise comparison conducted, our dual-stream C3D model consistently outperformed the SlowFast and dual-stream ConvNet models. Notably, the dual-stream C3D model achieved a Top-1 accuracy that was over 10% higher than that of the SlowFast model and surpassed the dual-stream ConvNet model by more than 5% across all viewpoint pairings.

The Robot-Front combination emerged as the most effective pairing, consistently yielding the highest results in both Top-1 and Top-5 accuracy metrics. The most impressive performance was observed with the dual-stream C3D model in the Robot-Front configuration, achieving a Top-1 accuracy of 71.06% and a Top-5 accuracy of 98.14%. These results underscore the effectiveness of the dual-stream C3D model, particularly when leveraging the Robot View in combination with other viewpoints.

V. Conclusion & Future Work

In conclusion, this study has substantiated the substantial benefits of integrating multi-view data into dual-stream models for Human Action Recognition (HAR), with a particular focus on enhancing the performance of the robot view. The inclusion of additional viewpoints as separate streams has been demonstrated to markedly improve the accuracy of the Robot View across various configurations. Notably, when the Robot View is positioned as the primary stream, it significantly enhances the model's ability to distinguish actions, indicating the effectiveness of this arrangement in leveraging multi-view data. Conversely, the study also reveals that the positioning of the Robot View as a secondary stream results in a relative decrease in accuracy, highlighting the importance of its role and placement within the model. Among the models compared, the dual-stream C3D model consistently outperforms alternatives like the SlowFast and dual-stream ConvNet models, evidencing its robustness and superior capability in handling multi-view data in HAR tasks. Furthermore, the combination of the Robot View with the Front View has been identified as the most effective, consistently yielding superior results in both Top-1 and Top-5 accuracy metrics. This optimal pairing, especially with the dual-stream C3D model, underscores the model's potential for practical applications in robot-enhanced environments.

Overall, the findings from this research provide significant insights into the application of multi-view data in robotic vision, emphasising the potential of the dual-stream C3D model in enhancing robot view performance. These insights are poised to inform future advancements in the field of robotic vision and HAR, contributing to the development of more sophisticated and accurate robotic systems across various domains.

In future research, there are two promising directions to explore for enhancing multi-view models in Human Action Recognition (HAR). The first involves leveraging temporal information from static views, aiming to unearth new layers of data that could improve the accuracy of multi-view models. This approach would focus on extracting dynamic elements from seemingly static images, potentially leading to more sophisticated recognition capabilities. The second area of exploration is the exclusive focus on the robot view dataset, developing deep learning models that rely solely on this perspective. Such an approach would concentrate on optimizing the unique insights provided by robotic perception, potentially leading to significant advancements in autonomous systems and a deeper understanding of robot-centric HAR. These two paths represent distinct but complementary strategies for advancing the field of robotic vision and action recognition.

References

[1] E. Broadbent, R. Stafford, and B. MacDonald, "Acceptance of healthcare robots for the older population: Review and future directions," *International journal of social robotics*, vol. 1, pp. 319–330, 2009.

[2] C. Breazeal, "Social interactions in hri: the robot view," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 2, pp. 181–186, 2004.

[3] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[4] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[5] D. Feil-Seifer and M. J. Mataric, "Defining socially assistive robotics," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, 2005, pp. 465–468.

[6] 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.

[7] D. C. Luvizon, H. Tabia, and D. Picard, "2d/3d pose estimation and action recognition using multitask deep learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5137–5146, 2018.

[8] J. Smith *et al.*, "Enhancing human action recognition in robot operations," *Journal of Robotic Systems*, vol. 39, no. 4, pp. 475–489, 2022.

[9] M. Jones and R. Kumar, "Adaptive learning models for human-robot interaction," in *Proceedings of the International Conference on Robotics and Automation*. IEEE, 2023.

[10] 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.

[11] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[13] E. Garcia and D. Lee, "Multi-perspective convolutional neural networks for improved human-robot interaction," *Robotics and Autonomous Systems*, vol. 140, p. 103665, 2022.

[14] W. Zhang *et al.*, "Integrated dual-stream networks for robotic vision systems," *Advanced Robotics*, vol. 37, no. 1, pp. 45–60, 2023.

[15] M. Bamorovat Abadi, M. R. Shahabian Alashti, P. Holthaus, C. Menon, and F. Amirabdollahian, "Multi-view fusion and feature extraction: Enhancing har for assistive robotics," *IEEE Access*, 2024.

[16] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.

[17] Y. Gu, X. Ye, W. Sheng, Y. Ou, and Y. Li, "Multiple stream deep learning model for human action recognition," *Image and Vision Computing*, vol. 93, p. 103818, 2020.

[18] L. Wang, Z. Wang, Y. Xiong, and Y. Qiao, "Cuhk&siat submission for thumos15 action recognition challenge," *THUMOS Action Recognition challenge*, pp. 1–3, 2015.

[19] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[21] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *arXiv preprint arXiv:1507.02159*, 2015.

[22] 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.

[23] Y. Wang, M. Long, J. Wang, and P. S. Yu, "Spatiotemporal pyramid network for video action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1529–1538.

[24] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4768–4777.

[25] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 2019, pp. 363–378.

[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.