MultiSensor-Home: Benchmark for Multi-modal Multi-view Action Recognition in Home Environments

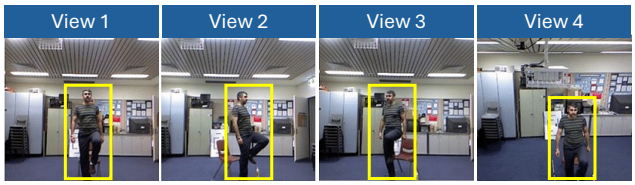
Trung Thanh Nguyen[1,2,a)] Yasutomo Kawanishi[2,3,b)] Vijay John[2,c)]
Takahiro Komamizu[3,1,d)] Ichiro Ide[1,3,e)]

Abstract

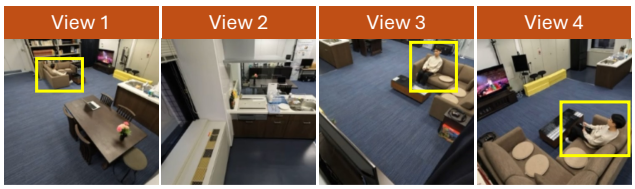
We present MultiSensor-Home, a benchmark for multi-modal multi-view action recognition in home environments. It consists of a dataset with 5,250 videos captured from five synchronized RGB+Audio sensors across two home environments, covering 17 distinct household activities. Each frame in these video sequences is manually annotated with fine-grained frame-level action labels, making it, to the best of our knowledge, the first densely annotated multi-view dataset for home activity recognition. To support fair and reproducible evaluation, we establish a comprehensive benchmark with standardized training and testing protocols. Evaluations are conducted under both joint-training and environment-specific settings to assess model generalization across diverse home layouts. We benchmark state-of-the-art methods on the proposed dataset, and the results highlight significant performance gaps across domains, especially in challenging scenes, providing valuable insights into the strengths and limitations of current approaches. This underscores the challenges and opportunities in multi-modal multi-view home action understanding and demonstrates the need for further advancements in the field.

1. Introduction

Action recognition is a fundamental problem in computer vision, with applications in video surveillance, assistive robotics, and smart home automation. Traditional single-view action recognition approaches [2], [3], [8] rely on a fixed viewpoint, which leads to occlusions, incomplete observations, and misclassifications. To address these limitations, multi-view action recognition has gained increasing attention by integrating multiple camera viewpoints, enabling more robust and accurate activity recognition. It exploits complementary viewpoints to mitigate occlusions and improve spatial reasoning. However, most existing datasets are designed for narrow-area settings, where multiple cameras capture the fully overlapping region to refine local action recognition (Figure 1(a)). While effective in controlled settings, such setups often fail to generalize to real-world home environments, where activities occur across multiple spatially separated locations. To address these challenges, we are focusing on the partially overlapping multi-view setting (Figure 1(b)), which involves action tracking across partially overlapping views, handling viewpoint variations, and performing efficient multi-modal data fusion.



(a) Fully overlapping multi-view setting.



(b) Partially overlapping multi-view setting (Ours).

Fig. 1: Configuration of multi-view settings.

Table 1 presents a summary of prominent existing benchmarks for multi-view action recognition. The NW-UCLA [9] and NTU RGB+D [4], [7] datasets are early efforts in multi-view action recognition, providing synchronized views but limiting their scope to narrow-area and short-duration videos. The Toyota Smarthome dataset [1] introduced RGB+D recordings in real home environments but lacks synchronized multi-view sequences. More recently, datasets featuring partially overlapping multi-view setting, such as MM-Office [11] and MM-Store [10], have emerged. These datasets incorporate distributed cameras capturing partially overlapping views, aligning more closely with real-world scenarios. However, they primarily provide video-level annotations, limiting their use in frame-level supervised learning

[1] Graduate School of Informatics, Nagoya University
[2] Guardian Robot Project, R-IH, RIKEN
[3] MDA Center, Nagoya University
[a)] nguyent@cs.is.i.nagoya-u.ac.jp
[b)] yasutomo.kawanishi@riken.jp
[c)] vijay.john@riken.jp
[d)] taka-coma@acm.org
[e)] ide@i.nagoya-u.ac.jp

Table 1: Comparison of the proposed MultiSensor-Home dataset with existing multi-modal multi-view action recognition datasets.

Settings	Dataset Name	Year	Modality	#Views	#Videos	Avg. Duration	#Classes	Resolution
Fully Overlapping Area	NW-UCLA [9]	2014	RGB+D	3	1,494	10 seconds	10	640 × 480
	NTU RGB+D [7]	2016	RGB+D	3	56,880	5 seconds	60	1,920 × 1,080
	NTU RGB+D 120 [4]	2019	RGB+D	3	114,480	5 seconds	120	1,920 × 1,080
	Toyota Smarthome [1]	2020	RGB+D	7	16,115	12 seconds	31	640 × 480
Partially Overlapping Area	MM-Office [11]	2022	RGB+Audio	4	1,760	60 seconds	12	2,560 × 1,440
	MM-Store [10]	2024	RGB+Audio	6	2,970	60 seconds	18	3,840 × 2,160
	MultiSensor-Home (Ours)	2025	RGB+Audio	5	5,250	70 seconds	17	4,000 × 3,000

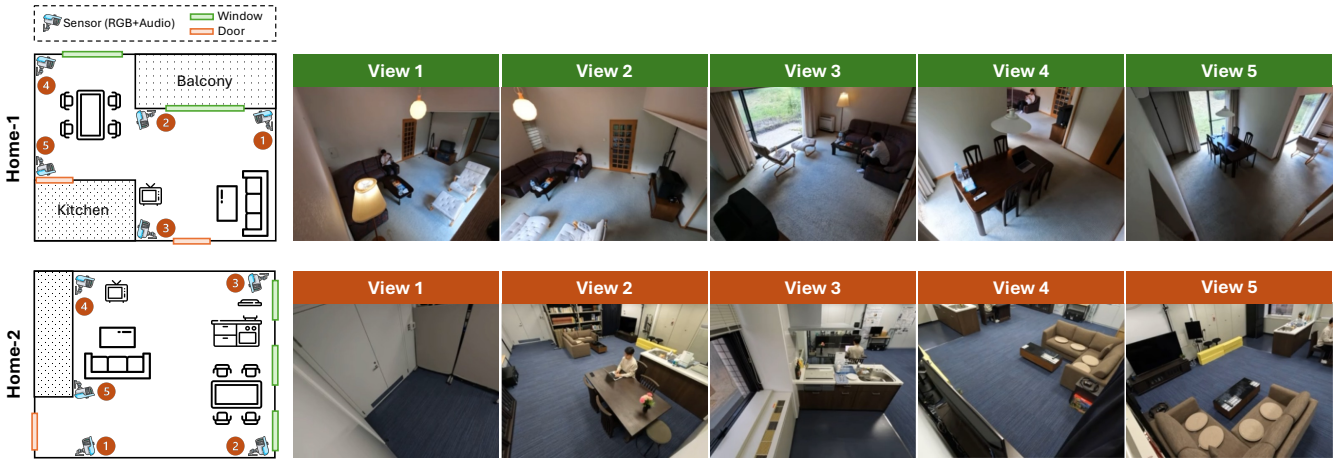


Fig. 2: Room layout illustrating the placement of multiple sensors in the proposed MultiSensor-Home dataset.

tasks. Furthermore, these datasets focus on controlled office and convenience store environments, making them less representative of home-based activities.

To address this limitation and build upon our previous study [6], we introduce the MultiSensor-Home benchmark for human action recognition in home environments, designed with the following key contributions:

- **Two home environments.** The MultiSensor-Home dataset is collected from two distinct home environments, incorporating variations in room layouts, lighting conditions, and background elements to improve generalization.
- **High-resolution frame-level annotations.** Each frame is carefully labeled with precise action annotations, ensuring fine-grained temporal analysis for action segmentation and recognition tasks.
- **Comprehensive benchmark.** We establish standardized training and evaluation protocols and assess state-of-the-art methods under both joint and environment-specific settings, providing a benchmark for multi-modal multi-view action recognition.

We believe that MultiSensor-Home should serve as a high-quality benchmark for advancing multi-modal multi-view action recognition, supporting research in wide-area activity understanding.

2. MultiSensor-Home Benchmark

MultiSensor-Home serves as a benchmark for multi-modal multi-view action recognition in real-world home environments, designed to facilitate the training and evaluation of recognition models. It features wide-area partially overlapping multi-view coverage, where synchronized RGB and Audio sensors capture activities across partially overlapping locations, enabling the study of action transitions in a wide-range environment. The dataset provides high-resolution and frame-level annotations, ensuring precise action segmentation. It incorporates diverse environmental conditions, covering two home environments, multiple times of day, and clothing variations to enhance model generalization. Additionally, the dataset contains untrimmed videos, allowing for action localization in continuous real-world scenarios with multiple activities per sequence.

2.1 Data Collection

The MultiSensor-Home dataset was collected in two distinct home environments, referred to as Home-1 and Home-2, to capture realistic indoor activities under diverse conditions. To ensure comprehensive spatio-temporal coverage, each home was equipped with five synchronized RGB cameras paired with an audio capture module. The cameras were strategically positioned in key home areas as illustrated in Figure 2. This configuration enables the capture of activities from partially overlapping viewpoints and provides consistent multi-view and multi-modal data.

Table 2 summarizes the dataset. All video recordings were conducted at a resolution of $4,000 \times 3,000$ pixels with a frame rate of 30 FPS. The dataset comprises 5,250 untrimmed videos, including 2,550 from Home-1 and 2,700 from Home-2, corresponding to 1,050 multi-view sequences. Each view contributes approximately 20.8 hours of video, re-

Table 2: MultiSensor-Home dataset overview and subset-specific statistics.

(a) General overview.

Attribute	Details
Sensor Settings	Wide-area, partially overlapping
Modality	RGB+Audio
Number of Views	5
Video Type	Untrimmed
Environment Type	Home (Indoor)
Time of Day	Morning, Afternoon, Evening
Number of People	Single
Clothing Variation	Multiple
Resolution	$4,000 \times 3,000$ pixels
Frame Rate	30 FPS
Synchronization	Yes
Annotations	Multi-view frame-level Labels
Class Distribution	Imbalanced

(b) Statistics for Home-1 and Home-2 environments.

Attribute	Home-1	Home-2
Total # of Videos	2,550	2,700
Total Duration	48 hours	56 hours
Average Video Duration	68 seconds	74 seconds
Average # of Actions in Video	3	3
# of Action Classes	16	16
Total # of Actions	1,334	1,171

sulting in a total duration of around 104 hours across all five synchronized views. The average length of the recordings in Home-1 is 68 seconds and 74 seconds in Home-2. Each video captures actions performed by a single individual in a natural home environment, with variations in illumination based on the time of day (morning, afternoon, and evening), clothing appearance, and ambient conditions.

The dataset was collected with informed consent, adhering to ethical standards on data privacy. No personally identifiable information is included. While anonymization is not applied to preserve data utility, all participants were made fully aware of the data usage and research intent.

2.2 Annotation

To ensure high-quality and consistent frame-level annotations, the MultiSensor-Home dataset follows a rigorous annotation process using Label Studio[*1], an open-source annotation tool. The annotation process is structured as follows:

- **Frame-level labeling.** Each frame is manually annotated with action labels, capturing the precise start and end timings of each action, facilitating fine-grained temporal analysis.
- **Multi-view synchronization.** The annotations are aligned across all camera views to ensure consistency, allowing accurate cross-view action tracking.
- **Multi-modal annotation.** Alongside visual annotations, the dataset incorporates synchronized audio data, enabling models to leverage both modalities for improved recognition.

[*1] Label Studio: Data labeling software. Open source software available from `https://github.com/HumanSignal/label-studio`. Accessed on 2025/06/05.

Table 3: Action classes in the proposed MultiSensor-Home dataset. "#Home-1" and "#Home-2" denote the number of occurrences in each environment.

Classes	Description	#Home-1	#Home-2
AdjustAC	Adjusting air conditioner	39	—
Clean	General cleaning activity	26	33
CleanVacuum	Cleaning with vacuum cleaner	48	45
OpenCurtain	Opening the curtain	38	35
CloseCurtain	Closing the curtain	39	35
Drink	Drinking water	51	59
Eat	Eating food	48	51
Enter	Entering the room	70	82
Exit	Exiting the room	88	82
ReadBook	Reading a book	64	79
Sitdown	Sitting down	247	142
Standup	Standing up	161	126
TurnOnLamp	Turning on the lamp	57	49
TurnOffLamp	Turning off the lamp	52	41
UseLaptop	Using a laptop computer	196	114
UsePhone	Using a phone	110	102
WatchTV	Watching the television	—	96
Total		1,334	1,171

Table 3 provides an overview of the 17 common household action classes and their occurrence frequencies within the dataset.

2.3 Benchmark

The MultiSensor-Home dataset[*2] includes all recorded videos, action annotations, and metadata to support further advancements in multi-modal multi-view action recognition. To ensure consistent evaluation, we also provide official train-test splits for the Home-1 and Home-2 environments, and the joint setup. This benchmark aims to facilitate the development of robust and generalizable models for understanding human activities in realistic home settings. This should serve as a high-quality benchmark that enables comprehensive evaluation and development of action recognition models, addressing key challenges in wide-area multi-modal multi-view action recognition.

3. Evaluation

3.1 Evaluation Metrics

To evaluate the performance of multi-label action recognition models on the MultiSensor-Home dataset, we adopt *sequence-level* and *frame-level* evaluation protocols. In the former setting, models are trained using sequence-wise action labels, while in the latter setting, frame-wise annotations are used for training.

To evaluate model performance, we report the following two mean Average Precision (mAP) metrics:

- mAP_C (macro-average mAP): Computes the average precision for each class and averages it across all classes, which is typically used for multi-label classification as it treats all classes equally.
- mAP_S (micro-average mAP): Computes AP over all samples, reflecting overall performance, which is commonly used in single-label classification settings.

3.2 Evaluation Methods

To establish benchmark performance on the proposed

[*2] It will be made publicly available for research purposes.

Table 4: Performance comparison of different methods when trained jointly on Home-1 and Home-2 environments, and evaluated separately on each environment. The best results are highlighted in **bold**.

Method	Sequence-level						Frame-level					
	Average		Home-1		Home-2		Average		Home-1		Home-2	
	mAP_C	mAP_S	mAP_C	mAP_S	mAP_C	mAP_S	mAP_C	mAP_S	mAP_C	mAP_S	mAP_C	mAP_S
MultiTrans [11]	66.67	80.51	58.01	78.14	75.34	82.87	74.50	86.54	64.99	80.71	84.00	92.37
MultiASL [5]	<u>68.70</u>	<u>85.87</u>	<u>58.69</u>	<u>80.79</u>	<u>78.70</u>	<u>90.94</u>	**82.52**	<u>89.68</u>	**74.68**	<u>85.24</u>	**90.35**	<u>94.13</u>
MultiTSF [6]	**73.26**	**90.68**	**64.87**	**88.54**	**81.65**	**92.84**	<u>82.38</u>	**93.11**	<u>74.60</u>	**90.10**	<u>90.16</u>	**96.11**

MultiSensor-Home dataset, we evaluate three representative multi-modal multi-view action recognition models. MultiTrans [11] employs view-wise temporal encoding followed by multi-head attention for feature fusion. MultiASL [5] introduces view-specific attention mechanisms with sequence-level supervision for handling audio-visual inputs. MultiTSF [6] integrates a Human Detection Module and fuses view-wise features using a unified temporal Transformer. All models are trained and evaluated under a consistent experimental setting, with results reported on the Home-1 and Home-2 environments, and a joint training setup.

3.3 Evaluation Results

We conduct two experimental setups to assess model performance on the MultiSensor-Home dataset. In the first setting, models are trained on the combined data from both Home-1 and Home-2 environments (joint training) and evaluated on each environment. In the second setting, models are trained and evaluated independently on each environment (home-specific training) to isolate environment-specific performance without cross-domain generalization. This evaluation setting enables the analysis of model performance under both shared and environment-specific conditions.

Joint Training. As shown in Table 4, MultiTSF [6] achieved the best performance across most evaluation levels, demonstrating the effectiveness of its unified spatio-temporal and multi-view fusion design. All models tended to perform better in the Home-2 environment than in Home-1, which suggests that Home-2 presents a more controlled environment with consistent lighting, less occlusion, and closer camera placement. In contrast, Home-1 contains more spatial variation, distant camera viewpoints, and natural lighting variability, making it more difficult for current models to generalize well across all views. These findings highlight the importance of robust model design for real-world and different environments.

Home-Specific Training. Table 5 presents results when models are trained and evaluated independently in each environment. This setup removes the benefit of cross-environment exposure and focuses on how models adapt to individual domains. The results show that the gap between Home-1 and Home-2 persists. For instance, MultiTSF improved significantly in Home-2, reaching 92.12% in mAP_C and 95.17% in mAP_S at the frame-level, but still lagged behind in Home-1. This reinforces the observation that

Table 5: Performance comparison of different methods on the Home-1 and Home-2 environments. All models are trained and evaluated independently on each dataset. The best and second-best results are highlighted in **bold** and <u>underlined</u> text, respectively.

Method	Sequence-level		Frame-level	
	mAP_C	mAP_S	mAP_C	mAP_S
Home-1				
MultiTrans [11]	<u>59.65</u>	<u>77.60</u>	61.40	78.07
MultiASL [5]	58.58	77.43	<u>73.81</u>	<u>85.38</u>
MultiTSF [6]	**64.48**	**87.91**	**76.12**	**91.45**
Home-2				
MultiTrans [11]	73.47	85.71	80.16	86.91
MultiASL [5]	<u>77.49</u>	<u>90.91</u>	<u>90.63</u>	<u>92.64</u>
MultiTSF [6]	**83.01**	**93.87**	**92.12**	**95.17**

Home-1 poses greater challenges due to occlusions, clutter, and distant views, while Home-2 provides a more favorable scenario for existing architectures.

These results demonstrate that the MultiSensor-Home dataset captures meaningful variations across home environments, making it a valuable and realistic benchmark for multi-modal multi-view action recognition. The findings also highlight the need for future models that can generalize more effectively across diverse spatial configurations and sensor placements.

4. Conclusion

We introduced MultiSensor-Home, a benchmark for multi-modal multi-view action recognition in realistic home environments. The dataset includes 5,250 untrimmed videos across two home environments, captured with five synchronized RGB and Audio sensors and annotated at the frame-level. Experimental results across joint and home-specific training settings highlight the challenges posed by diverse home layouts, camera placements, and lighting conditions. Current methods show performance gaps, particularly in more complex environments.

For future work, we plan to develop a federated learning framework for model training across homes without centralizing data. This direction addresses privacy concerns inherent in home environments and supports preserving privacy.

Acknowledgments

This work was partly supported by JSPS KAKENHI JP21H03519 and JP24H00733.

References

[1] Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F. and Francesca, G.: Toyota Smarthome: Real-world activities of daily living, *Proceedings of the 17th IEEE/CVF International Conference on Computer Vision*, pp. 833–842 (2019).

[2] Gupta, P., Thatipelli, A., Aggarwal, A., Maheshwari, S., Trivedi, N., Das, S. and Sarvadevabhatla, R. K.: Quo vadis, skeleton action recognition?, *International Journal of Computer Vision*, Vol. 129, No. 7, pp. 2097–2112 (2021).

[3] Kong, Y. and Fu, Y.: Human action recognition and prediction: A survey, *International Journal of Computer Vision*, Vol. 130, No. 5, pp. 1366–1401 (2022).

[4] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y. and Kot, A. C.: NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 10, pp. 2684–2701 (2019).

[5] Nguyen, T. T., Kawanishi, Y., Komamizu, T. and Ide, I.: Action selection learning for multilabel multiview action recognition, *Proceedings of the 2024 ACM Multimedia Asia Conference*, pp. 1–7 (2024).

[6] Nguyen, T. T., Kawanishi, Y., Vijay, J., Komamizu, T. and Ide, I.: MultiSensor-Home: A wide-area multi-modal multiview dataset for action recognition and Transformer-based sensor fusion, *Proceedings of the 19th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–10 (2025).

[7] Shahroudy, A., Liu, J., Ng, T.-T. and Wang, G.: NTU RGB+D: A large scale dataset for 3D human activity analysis, *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019 (2016).

[8] Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G. and Liu, J.: Human action recognition from various data modalities: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 3, pp. 3200–3225 (2022).

[9] Wang, J., Nie, X., Xia, Y., Wu, Y. and Zhu, S.-C.: Cross-view action modeling, learning and recognition, *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2649–2656 (2014).

[10] Yasuda, M., Harada, N., Ohishi, Y., Saito, S., Nakayama, A. and Ono, N.: Guided masked self-distillation modeling for distributed multimedia sensor event analysis, *Computing Research Repository arXiv Preprints,* arXiv:2404.08264, pp. 1–13 (2024).

[11] Yasuda, M., Ohishi, Y., Saito, S. and Harado, N.: Multi-view and multi-modal event detection utilizing Transformer-based multi-sensor fusion, *Proceedings of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4638–4642 (2022).