scientific data

OPEN

DATA DESCRIPTOR

MCV-Intention: A Multimodalities and Cross-View Dataset for Human Assembly Intention Recognition

Dongxu Ma[1], Chao Zhang[1,2] ✉, Qingfeng Xu[1] & Guanghui Zhou[1,2] ✉

In recent years, Industry 5.0 has emphasized human-centric intelligent manufacturing, positioning human-robot collaboration as a cornerstone for achieving mass customization. Consequently, enabling robots to perceive human state has become critical for efficient and safe human-robot collaborative assembly. However, current vision-based methods for assembly intention recognition face challenges such as limited dataset modalities, difficulties in reflecting real-world assembly processes, and inconsistent annotation workflows. To address these issues, this paper introduces the MCV-Intention dataset—a multimodalities, cross-view dataset designed for assembly scene understanding. Collected from 15 subjects, each assembly sequence encompasses six modalities and two views, capturing data from operators both before and after training. In this paper, we first outline the dataset collection process, detailing the hardware and software systems as well as the assembly objects involved. Subsequently, we present a comprehensive annotation protocol for assembly intention recognition and analyze the dataset from the viewpoints of structure and distribution. Finally, we conducted a series of benchmark experiments using state-of-the-art algorithms to establish baselines for future researches.

Background & Summary

Human robot collaborative (HRC) paradigm exhibits great potential pathway to achieve mass customization[1], which takes full advantage of precision and repeatability of robot, and cognition, flexibility of human operators to achieve ergonomic working conditions with better productivity. In this context, the efficient recognition of human assembly intentions by robots is crucial for ensuring both the efficiency and safety of collaborative assembly processes[2].

Currently, vision-based human assembly intention recognition approaches have gained more and more attentions, such as single modal-based approaches[3], dual modal-fusion approaches[4], triple modal-fusion approaches[5]. However, these works unanimously emphasized the absence of sufficient datasets within the industrial context. And most approaches were verified with custom datasets, which is unfavorable for the investigation of high-performance, cross-scenes application aimed at recognizing operator intentions. While existing datasets have been developed, they suffer from three critical shortcomings[6,7]: (1) limited modalities, impeding full exploitation of spatiotemporal information; (2) confinement to predefined assembly cases, overlooking stochastic operator behaviors; (3) non-standardized annotation frameworks and protocols, obstructing systematic data utilization. Table 1 presents an overview of several datasets currently available for assembly scenarios. Hence, in this paper, we propose a new human assembly intention recognition dataset MCV-Intention that is collected within real industrial scenario with six modalities and two views using custom data acquisition software developed by our research team. The main contribution of this paper are as follows:

- Leveraging our custom data acquisition software, a multimodalities dataset encompassing six modalities and two views are collected (MCV-Intention), with small satellite serving as the assembly object. The entire data collection process is conducted by fifteen subjects. This dataset comprehensively simulates real-world assembly scenarios, incorporating both unfamiliar assembly and post-training proficient assembly.
- An annotation protocol encompassing comprehensive details is established.

[1]School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. [2]State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, 710054, China. ✉e-mail: superzc@xjtu.edu.cn; ghzhou@mail.xjtu.edu.cn

Cit	Provided modal						
	RGB	Sk	Depth	OF	PC	IR	Mask
HRCA[5]	✓	✓	✓	×	×	×	×
WB[6]	✓	×	×	×	×	×	×
InHard[7]	✓	✓	✓	×	×	×	×
HA4M[21]	✓	✓	✓	×	✓	✓	×
HA-ViD[22]	✓	✓	×	×	×	×	×
Ikea[23]	✓	✓	✓	×	×	×	×
Brio[24]	✓	×	×	×	×	×	×
Attach[25]	✓	✓	×	×	×	×	×
Meccano[26]	✓	×	✓	×	×	×	×

Table 1. Existing dataset for human assembly intention recognition.

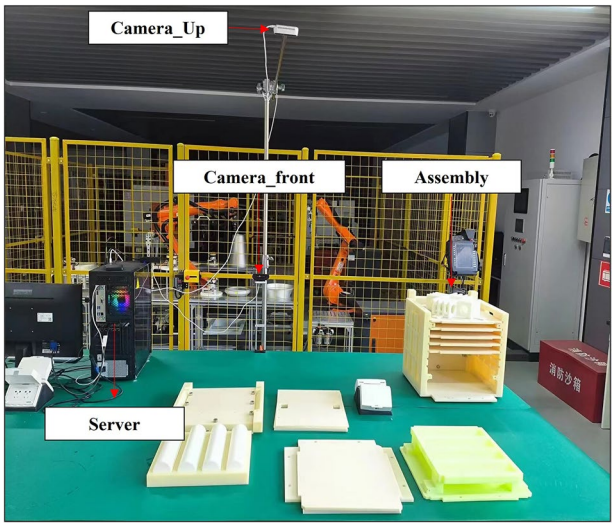


Fig. 1 Hardware environment for data acquisition.

- A comprehensive suite of benchmark experiments is conducted using state-of-the-art algorithms founded on convolutional techniques and attention mechanisms. It can be viewed as the baseline for research community related this topic.

Methods

Dataset acquisition protocol. Most existing datasets are collected with predefined rules, where every assembly action is constrained, and no errors are permitted during the procedure. However, there is a significant likelihood that operators will commit errors or unexpected behaviors during task execution, such as pause, disassembly, wrong usage of tools, etc. This may result in the human assembly intention model exhibiting excellent performance during the training phase, yet demonstrating suboptimal efficacy in practical applications. Therefore, to remedy this gap, we propose MCV-Intention, a multimodalities and cross-view human assembly intention recognition dataset with natural human behaviors. The details of dataset acquisition processes are as follows:

(1) The subjects are required to perform the assembly process twice, namely pre-training and after post-training.
(2) Before training, the subjects are not aware of the assembly process and only gained information with assembly documents, where some abnormal actions will occur.
(3) After training, the subjects are required to perform assembly tasks with meticulous guidance to ensure the successful and normal completion of the assembly procedure.

Dataset acquisition environment. Here, we build the dataset acquisition environment from the viewpoint of hardware and software environments, as shown in Fig. 1.

Hardware environment. Hardware environment is an execution cell where assembly process is conducted. It consists of assembled objects (small satellite), tools, a top-down camera, a front-camera, and a server, as shown

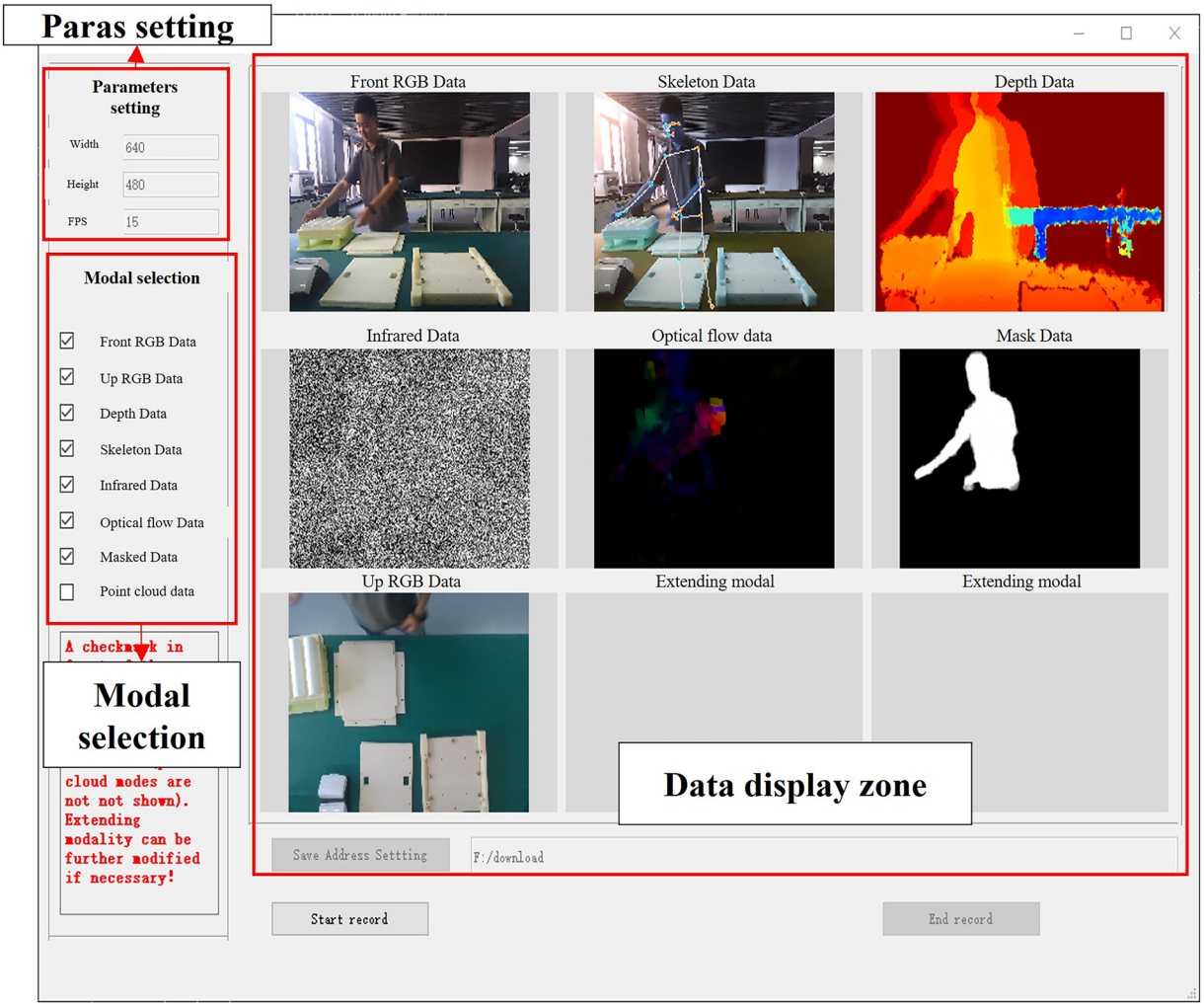


Fig. 2 Custom data acquisition software.

in Fig. 1. Here, top-down camera is responsible for collecting top-down view of assembly process (the last display zone in Fig. 2). And the rest of modalities are captured by front-facing camera (the 1–6 display zone in Fig. 2).

Software environment. The software is developed by our research team using PyQT, which is mainly used to collect different modalities and views data during assembly process. The parameter setting area is used to set width and height of captured images, and frames per second (*fps*). In addition, modal selection can be done in modal selection zone according user's requirements. All captured data are virtualized in data display zone, as shown in Fig. 2. Here, front and top-down RGB modal data are directly captured with cameras. Skeleton and mask modal are obtained with mediapipe of google. Optical flow modal is calculated by Lucas-Kanade method[8]. Depth and infrared modal are acquired with camera SDK.

Assembled object. Most existing human assembly action recognition datasets are compiled through the use of significantly simplified assemblies, such as toy, simplified gear box, etc., thereby failing to adequately represent the complexities of real-world assembly operations. Therefore, we use a small satellite as assembled object, as shown in Fig. 3. It consists of 9 major parts (including radiator, inner plane, holder, solar plane, battery holder, chassis, battery, play load, and top plane), 23 submodules, and 100 screws. The small satellite has overall dimensions of 400 mm × 400 mm × 400 mm, weighs 35 kg, and was fabricated using additive manufacturing (3D printing) upon design completion. The entire assembly process typically requires approximately thirty minutes to complete when performed by a skilled worker. The assembly process of small satellite is depicted in Fig. 4. The assembly steps can be specified as follows: (1) battery module assembly, (2) left plane assembly, (3) right plane assembly, (4) left solar plane assembly, (5) right solar plane assembly, (6) front solar plane assembly, (7) back solar plane assembly, (8) heat dissipation assembly, (9) finally assembly. In addition, the detailed assembly documents and process can be found from our dataset website.

Annotation process. The purpose of dataset annotation is to differentiate various actions. The annotation process involves manual frame-by-frame labeling, followed by verification by three additional researchers to

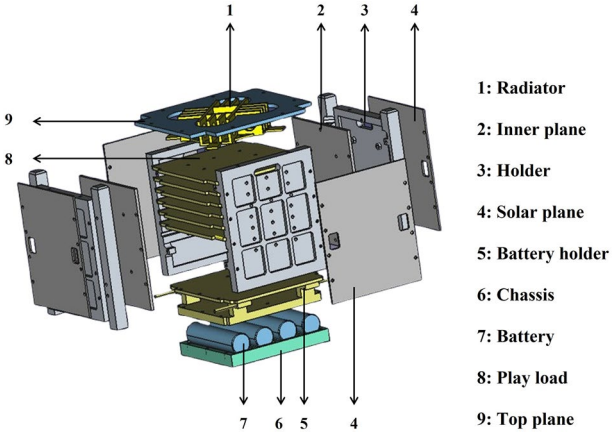


Fig. 3 The small satellite. Constructed by opensource software FreeCAD. Original model can be download here.

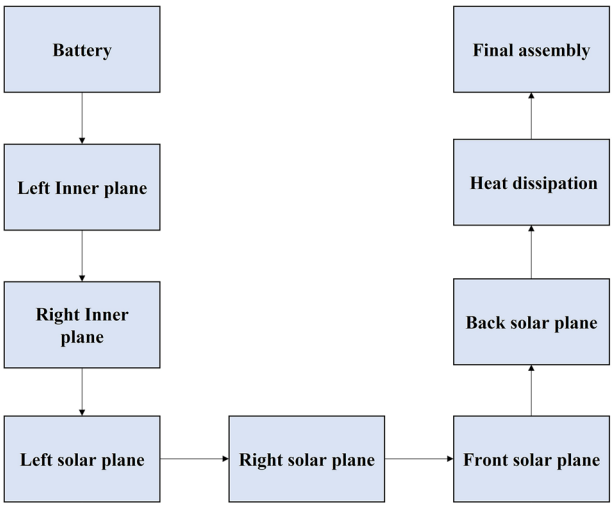


Fig. 4 Assembly procedures of small satellite.

ensure accuracy. Examples of dataset annotation are presented in Table 2. Different actions are denoted by distinct IDs, encompassing both normal and abnormal actions. The annotation example is shown in Table 2.

Ethical approval. This research was conducted under protocols reviewed and approved by the Institutional Review Board (IRB) of The First Affiliated Hospital of Xi'an Jiaotong University (IRB No: XJTU1AF2025LSYY-036). Participants were recruited on a voluntary basis from the university community. We established specific inclusion and exclusion criteria to ensure participant suitability. Inclusion criteria required healthy adults with the normal vision, cognitive function, and motor skills necessary for the assembly task. Individuals with any physical or cognitive impairments that could affect task performance, or those with prior familiarity with the assembly process, were excluded. Prior to their involvement, all participants provided written informed consent. The consent form explicitly detailed the study's procedures, potential minimal risks (including minor physical fatigue and discomfort from being filmed), and the data handling policy. Crucially, participants were informed and consented that their fully anonymized data, including video recordings, would be made publicly available for research purposes through scientific data repositories like Zenodo. The research team adhered to all approved guidelines for data collection, cleaning, storage, and dissemination.

Data Records

To minimize data redundancy, we set height and width as 480 and 640 in pixels while *fps* is 15. There are mainly three steps about data collection: (1) For pre-training phase, all 15 subjects, initially unfamiliar with assembly procedures, were provided with process documentation to guide their attempts. Their lack of familiarity led to various assembly issues, capturing their pre-training state data; (2) For training phase, an experienced assembly worker trains the subject until they are familiar with the entire assembly process and can complete the assembly

Action	Normal ID	Abnormal ID	Action description
Battery module	0	1	Four batteries should be inserted to cell bucket, and a header cover installed subsequently.
Left inner plane	2	3	Install the inner plane into the left main framework.
Right inner plane	4	5	Install the inner plane into the right main framework.
Left solar plane	6	7	Install the solar plane into the left main framework.
Right solar plane	8	9	Install the solar plane into the right main framework.
Front solar plane	10	11	Install the solar plane into the front main framework.
Back solar plane	12	13	Install the solar plane into the back main framework.
Heat dissipation	14	15	Install heat dissipation into payload.
Final assembly	16	17	Assemble all subsystems of the satellite into a complete unit.

Table 2. dataset annotation details.

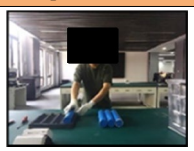
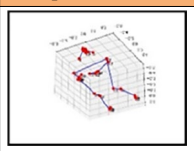
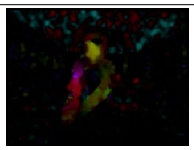
View	Modal	Example	Format	View	Modal	Example	Format
Front view	RGB		.jpg	Front view	Skeleton		.txt
Front view	Depth		.jpg	Front view	IR		.jpg
Front view	OF		.jpg	Front view	Mask		.jpg
Top-down view	RGB		.jpg				

Table 3. Dataset example. OF denotes optical flow, IR is infrared.

without referring to the manual; (3) For post-training phase, after becoming familiar with the process, subjects performed the post-training assembly, where no-error process is required for each assembly step. If any error occurs, the subjects will be asked to perform assembly step again. Data collection process is conducted on a per-assembly-step basis. Finally, a total size of 32 GB of MCV-Intention dataset is collected following the protocol. The dataset has been organized and uploaded to zenodo, which can be obtained from zenodo database[9]. The dataset examples are shown in Table 3. There is total six modalities (RGB, depth, skeleton, optical flow, infrared, mask) and two views (front view and top-down view), where skeleton modal is stored in *.txt* file format, primarily containing three-dimensional coordinates (*x, y, z*) for 33 human skeletal joints and other modalities are saved with *.jpg* format. The dataset is organized as shown in Fig. 5.

Technical Validation

Dataset analysis. The dataset distribution is shown in Fig. 6, where abnormal and normal means pre-training and post-training. Here, all abnormal and normal frames are calculated separately. The duration for completing assembly tasks varies among subjects, reflecting their diverse levels of experience with such operations. Moreover, during pre-training, all subjects committed errors, such as arbitrary pauses, incorrect installations, and omissions, thereby extending the time required for assembly tasks before training, as depicted in Fig. 6(a). In addition, the time expenditure for each step varies as a result of differences in the complexity of the installation processes. As shown in Fig. 6(b), the installation of the battery required the least amount of time, whereas the installation of the heat dissipation consumed the most time. Here, battery assembly procedure is intuitive and simple in design while heat dissipation installation requires specialized tools under narrow space.

Here, to evaluate the efficiency and quality of MCV-Intention, we reproduce the 8 outstanding algorithms for human action recognition tasks based on convolution and attention mechanism, namely CorrNet[10], CSN[11], ResNet3D[12], SlowFast[13], ViViT[14], MViT[15], AcT[16], and SR2M[4]. Due to the informational efficacy of RGB and

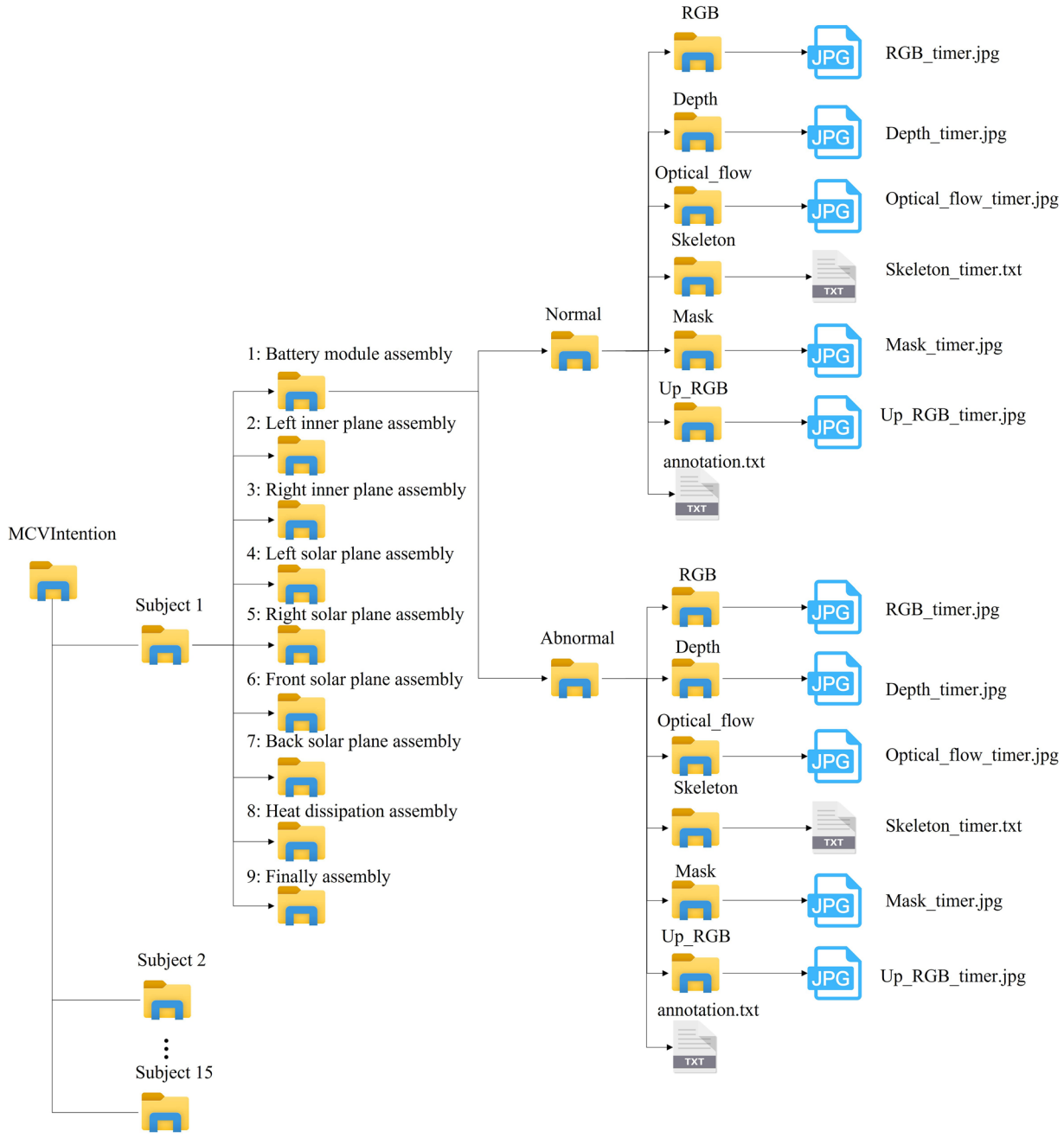
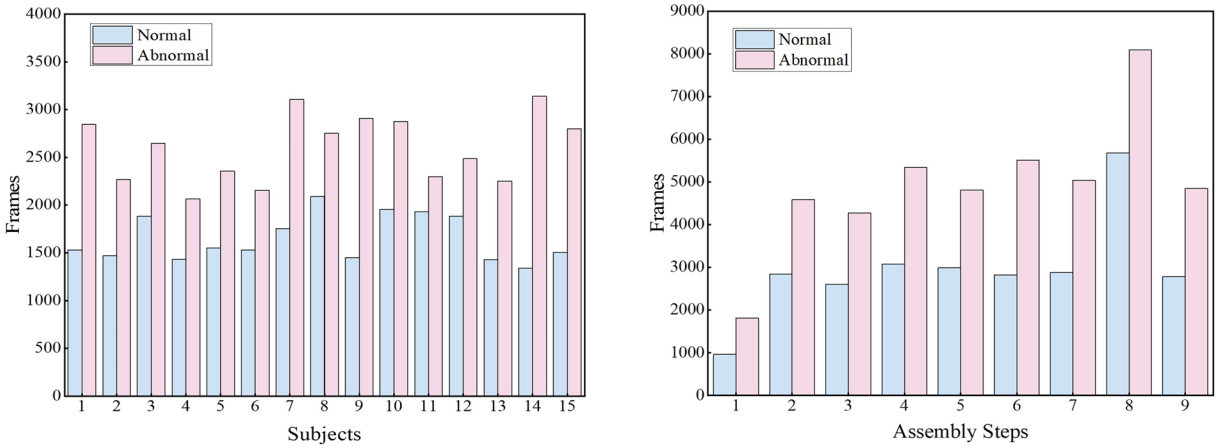


Fig. 5 Dataset structure. Subject *i* means different volunteers. 1–9 denotes assembly steps according to the order.

skeleton modalities, they are frequently employed as source modalities for assembly intention recognition[17]. Therefore, the RGB and skeleton modal are used to demonstrate the efficiency of MCV-Intention dataset. The detailed information of above-mentioned models are as follows.

CorrNet short for correlation networks, leverages a learnable correlation operator to establish frame-to-frame correspondences across convolutional feature maps in various network layers. CSN is a 3D channel-separated network, where all convolutional operations are separated into either pointwise $1 \times 1 \times 1$ or depthwise $3 \times 3 \times 3$ convolutions. ResNet3D is proposed for human assembly recognition based on context information. It contains residual convolutional neural network with 34 layers and a long short-term memory recurrent neural network. SlowFast contains slow pathway and fast pathway for feature extraction. Here, slow pathway extracts spatial semantics based on low frame rate, while fast pathway is utilized to capture motion with fine temporal resolution based on high frame rate. ViViT refers to video vision transformer, which is a pure-transformer architecture-based video classification network. MViT is multiview transformers network, which is also developed with transformer architecture. It uses separate transformer encoder to represent different views of one



(a) Subjects and frames (b) Assembly steps and frames

Fig. 6 Total frames of normal and abnormal assembly process (examples with RGB modal).

Setting parameters	value
Maximum epoch	700
Learning rate	1e-4
Input frame size	112×112
Input frame number	16
Batch size	16

Table 4. The Setting of Benchmark Experiments.

input video with a lateral connection to integrate features across different views. AcT is action transformer, a simple, fully, self-attentional architecture for action recognition based on skeleton modal. SR2M denotes skeleton-RGB integrated network for human assembly action prediction. Here, 3D Resnet model[18] is used to extract features of RGB modal, while multi-scale graph convolution model[19] is responsible for feature extraction of skeleton modal.

And the training process setting is depicted in Table 4.

We conduct benchmark experiments on two highly used modalities (RGB and skeleton). Here, model accuracy refers to the video classification accuracy, defined as the proportion of correctly classified samples for a specific action category relative to the total number of samples in that category. As shown in Table 5, for single-modality input with normal schema, the highest accuracy achieved is only 90.25% with SlowFast, while attention-based models generally exhibit mediocre performance. For instance, ViViT achieves an accuracy of merely 65.24%, and MViT reaches 81.58%. Similarly, the skeleton modal-based AcT model records a notably lower accuracy of 60.00%, potentially attributable to the limited contextual information it provides, which hinders its ability to recognize complex assembly processes. In contrast, the multi-modalities-based SR2M model demonstrates a superior accuracy of 91.20%. On the other hand, for abnormal assembly actions, their inherent high uncertainty poses significant challenges for algorithms, particularly for attention-based algorithms. For instance, ViViT achieves an accuracy of only 31.22%, while AcT attains a mere 30.14%. In contrast, other convolution-based algorithms outperform their attention-based counterparts, with SlowFast achieving an accuracy of 66.12%. Furthermore, the multimodalities fusion algorithm SR2M demonstrates a certain advantage in handling uncertainty, reaching an accuracy of 70.23%.

It can be easily concluded from benchmark experiments that those state-of-the-art algorithms perform suboptimal performance for abnormal assembly process with maximum accuracy is 70.23%, while maximum accuracy for normal assembly process is over 90%. It falls short of the standards required for practical application. In contrast to a normal assembly process, an abnormal assembly process encompasses a variety of aberrant behaviors exhibited by operators. These behaviors include repeated adjustments to unfamiliar assembly tasks, the use of inappropriate tools, errors in assembly followed by disassembly and reassembly, and intermittent pauses, etc., as shown in Fig. 7. This significantly increases the uncertainty of the system. The distinct habits of individual assemblers contribute to an anomalous distribution within the dataset, which may constitute the primary reason for the reduced accuracy observed in the anomalous dataset. However, real-world assembly scenarios more closely resemble abnormal conditions. Therefore, more efficient algorithms are needed.

In addition, compared to models based on convolutional operations, models leveraging attention mechanisms exhibit inferior performance, potentially attributable to their limited inductive bias capability. Furthermore, owing to the complexity of scenes, these algorithms may fail to adequately extract and integrate

Model	Modal		Normal	Abnormal
	RGB	Skeleton		
CorrNet	✓	×	76.00%	45.34%
CSN	✓	×	61.25%	39.76%
ResNet3D	✓	×	84.20%	58.60%
SlowFast	✓	×	90.25%	66.12%
ViViT	✓	×	65.24%	31.22%
MViT	✓	×	81.58%	53.21%
AcT	×	✓	60.00%	30.14%
SR2M	✓	✓	**91.20%**	**70.23%**

Table 5. Benchmark experiment of MCV-Intention on the state-of-the-art methods.

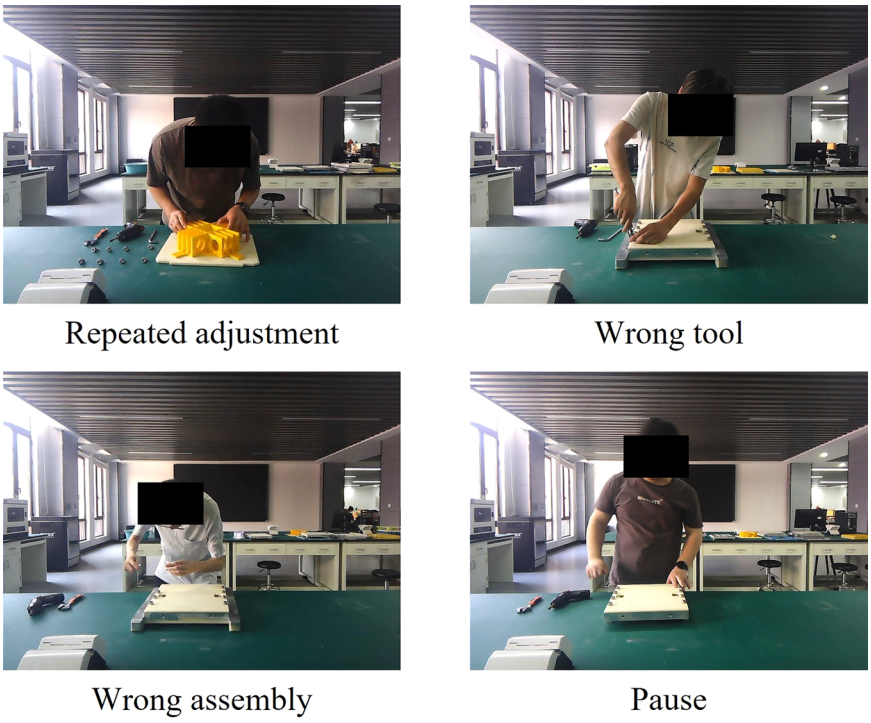


Fig. 7 Abnormal behaviors of operators.

contextual information, resulting in diminished system robustness. Additionally, multimodalities fusion demonstrates superior performance, suggesting that the alignment of different modalities could prove more effective and decrease the uncertainties in enhancing overall system efficacy[20]. Therefore, future research on assembly intention recognition should not solely focus on accuracy. Instead, it must fully leverage multimodalities datasets to effectively extract temporal and spatial information in different resolutions, while quantifying the uncertainty of human operators. This approach will enhance the overall robustness of HRC assembly systems.

Data availability

The dataset can be found in https://doi.org/10.5281/zenodo.15083791. The benchmark dataset of experiments can be downloaded from google driver with link https://drive.google.com/file/d/1dtnAbYm38ZpeDJG2j7xyCsghp4JH8Kjg/view?usp=sharing.

Code availability

The resulting dataset, totaling 32 GB, along with the collection software, annotation tool, algorithms code, is made available at https://github.com/mdx-box/human-assembly-intention-recognition/tree/master.

Received: 31 March 2025; Accepted: 26 September 2025;
Published online: 13 November 2025

References

1. Zhang, C. *et al*. Towards new-generation human-centric smart manufacturing in Industry 5.0: A systematic review. *Adv. Eng. Inform.* **57**, 102121, https://doi.org/10.1016/j.aei.2023.102121 (2023).
2. Jones, J. D. *et al*. Fine-grained activity recognition for assembly videos. *IEEE Robot. Autom. Lett.* **6**, 3728–3735, https://doi.org/10.1109/LRA.2021.3064149 (2021).
3. Zhang, J., Wang, P. & Gao, R. X. Hybrid machine learning for human action recognition and prediction in assembly. *Robot. Comput.-Integr. Manuf.* **72**, 102184, https://doi.org/10.1016/j.rcim.2021.102184 (2021).
4. Zhang, Y. *et al*. Skeleton-RGB integrated highly similar human action prediction in human–robot collaborative assembly. *Robot. Comput.-Integr. Manuf.* **86**, 102659, https://doi.org/10.1016/j.rcim.2023.102659 (2024).
5. Wang, T., Liu, Z., Wang, L., Li, M. & Wang, X. V. Data-efficient multimodal human action recognition for proactive human–robot collaborative assembly: A cross-domain few-shot learning approach. *Robot. Comput.-Integr. Manuf.* **89**, 102785, https://doi.org/10.1016/j.rcim.2024.102785 (2024).
6. Zhang, J., Byvshev, P. & Xiao, Y. iA video dataset of a wooden box assembly process: dataset, in *Proc. Third Workshop Data: Acquis. Anal*, pp. 35–39, https://doi.org/10.1145/3419016.3431492 (2020).
7. Dallel, M., Havard, V., Baudry, D. & Savatier, X. Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics, in *2020 IEEE Int. Conf. Hum.-Mach. Syst*. IEEE, pp. 1–6, https://doi.org/10.1109/ICHMS49158.2020.9209531 (2020).
8. Sánchez, J., Salgado, A. & Monzón, N. Computing inverse optical flow. *Pattern Recognit. Lett.* **52**, 32–39, https://doi.org/10.1016/j.patrec.2014.09.009 (2015).
9. Ma, D., Zhang, C., Xu, Q. & Zhou, G. MCV-Intention: A Multimodalities and Cross-View Dataset for Human Assembly Intention Recognition. *Zenodo* https://doi.org/10.5281/zenodo.15083791 (2025).
10. Wang, H., Tran, D., Torresani, L. & Feiszli, M. Video modeling with correlation networks, in *Proceedings of the IEEE/CVF conf. comput. vis. pattern recognit*, pp. 352–361, https://arxiv.org/abs/1906.03349 (2020).
11. Tran, D., Wang, H., Torresani, L. & Feiszli, M. Video classification with channel-separated convolutional networks, in *Proc. IEEE/CVF int. conf. comput. vis*. pp. 5552–5561 (2019).
12. Moutinho, D., Rocha, L. F., Costa, C. M., Teixeira, L. F. & Veiga, G. Deep learning-based human action recognition to leverage context awareness in collaborative assembly. *Robot. Comput.-Integr. Manuf.* **80**, 102449, https://doi.org/10.1016/j.rcim.2022.102449 (2023).
13. Feichtenhofer, C., Fan, H., Malik, J. & He, K. Slowfast networks for video recognition, in *Proc. IEEE/CVF int. conf. comput. vis*. pp. 6202–6211, https://doi.org/10.48550/arXiv.1812.03982 (2019).
14. Arnab, A. *et al*. Vivit: A video vision transformer, in *Proc. IEEE/CVF int. conf. comput. vis*. pp. 6836–6846, https://doi.org/10.48550/arXiv.2103.15691 (2021).
15. Yan, S. *et al*. Multiview transformers for video recognition, in *Proc. IEEE/CVF conf. comput. vis. pattern recognit*. pp. 3333–3343, https://doi.org/10.48550/arXiv.2201.04288 (2022).
16. Mazzia, V., Angarano, S., Salvetti, F., Angelini, F. & Chiaberge, M. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognit.* **124**, 108487, https://doi.org/10.1016/j.patcog.2021.108487 (2022).
17. Ma, D. *et al*. A Systematic Review on Vision-Based Proactive Human Assembly Intention Recognition for Human-Centric Smart Manufacturing in Industry 5.0. *IEEE Internet Things J*. 1-1, https://doi.org/10.1109/JIOT.2025.3570510 (2025).
18. Tran, D. *et al*. A closer look atspatiotemporal convolutions for action recognition, in: *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit*. IEEE, Salt Lake City, USA, pp. 6450–6459, https://doi.org/10.1109/cvpr.2018.00675 (2018).
19. Liu, Z. *et al*. Disentangling and unifying graph convolutions for skeleton-based action recognition, in: *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit*. IEEE, Seattle, USA, pp. 140–149, https://doi.org/10.1109/cvpr42600.2020.00022 (2020).
20. Park, D., Hoshi, Y. & Kemp, C. C. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robot. Autom. Lett.* **3**, 1544–1551, https://doi.org/10.1109/LRA.2018.2801475 (2018).
21. Cicirelli, G. *et al*. The HA4M dataset: Multi-Modal Monitoring of an assembly task for Human Action recognition in Manufacturing. *Sci. Data* **9**, 745, https://doi.org/10.1038/s41597-022-01843-z (2022).
22. Zheng, H., Lee, R. & Lu, Y. Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding. *Advances in Neural Information Processing Systems* **36**, 67069–67081, https://doi.org/10.48550/arXiv.2307.05721 (2023).
23. Ben-Shabat, Y. *et al*. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose, in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis*, pp. 847–859, https://doi.org/10.48550/arXiv.2007.00394 (2021).
24. Moriwaki, K., Nakano, G. & Inoshita, T. The brio-ta dataset: Understanding anomalous assembly process in manufacturing, in *2022 IEEE Int. Conf. Image Process*. IEEE, pp. 1991–1995, https://doi.org/10.1109/ICIP46576.2022.9897369 (2022).
25. Aganian, D., Stephan, B., Eisenbach, M., Stretz, C. & Gross, H.-M. ATTACH dataset: Annotated two-handed assembly actions for human action understanding, in *2023 IEEE Int. Conf. Robot. Autom*. IEEE, pp. 11367–11373, https://doi.org/10.1109/ICRA48891.2023.10160633 (2023).
26. Ragusa, F., Furnari, A. & Farinella, G. M. Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. *Comput. Vis. Image Underst.* **235**, 103764, https://doi.org/10.1016/j.cviu.2023.103764 (2023).

Acknowledgements

The authors would like to thank all subjects who participated in the data collection experiments. In addition, the authors thank the foundation from National Natural Science Foundation of China under Grant number 52541501, 52375511 and 52475534.

Author contributions

Dongxu Ma and Qingfeng Xu created the experimental protocol, wrote the code for data acquisition and annotation, manipulation, and visualization. Chao Zhang and Guanghui Zhou planned and supervised the projects and reviewed the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.Z. or G.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.