Industrial Datasets for Multi-Modal Monitoring of an Assembly Task for Human Action Recognition and Segmentation

Laura Romeo[1,*], Annaclaudia Bono[1,2], Grazia Cicirelli[1] and Tiziana D'Orazio[1]

[1]Institute of Intelligent Industrial Systems and Technologies for Advanced Manufacturing (STIIMA), National Research Council (CNR), Bari, Italy
[2]Department of Electrical and Information Engineering (DEI), Polytechnic of Bari, Bari, Italy

Abstract

With the rapid evolution of advanced industrial systems exploiting deep learning techniques, the availability of multimodal and heterogeneous datasets of operators working in industrial scenarios is essential. Such datasets allow in-depth studies for accurate segmentation and recognition of the actions of operators working alongside collaborative robots. Using multimodal information guarantees the capture of relevant features to analyze human movements properly. This paper presents our recent research activity on the development of two datasets representing human operators performing assembly tasks in industrial contexts. The dataset for Human Action Multi-Modal Monitoring in Manufacturing (HA4M) is a collection of multimodal data recorded using a Microsoft Azure Kinect camera observing 41 subjects while performing 12 actions to assemble an Epicyclic Gear Train (EGT). The dataset for Human-Cobot Collaboration for Action Recognition in Manufacturing Assembly (HARMA) focuses on the interaction between 27 subjects and a collaborative robot while assembling the EGT in 7 actions. In this case, the acquisition setup consisted of two Microsoft Azure Kinect cameras. Both datasets were collected in controlled laboratories. To prove the validity of the HA4M and HARMA datasets, state-of-the-art temporal action segmentation models, i.e. MS-TCN++ and ASFormer, were trained using both skeletal and video features. The results successfully prove the effectiveness of the presented datasets in segmenting human actions in industrial contexts.

Keywords

Image processing, Assembly Datasets, Action Segmentation, Action Recognition, Manufacturing

1. Introduction

In Industry 5.0, the interaction between humans and collaborative robots (cobots) is becoming more and more important for manufacturing processes [1]. Cobots represent a shift in robotic technology. Traditional robots typically operate in confined work cells or dedicated spaces having predefined and automated tasks. Unlike traditional robots, cobots operate in environments where they can interact directly with human workers to solve tasks that require a combination of human cognition and robot strength and repeatability.

In manufacturing processes, human action recognition and segmentation are crucial for many reasons: to promote human-robot cooperation [2]; to assist operators [3]; to support employee training [4, 5]; to increase productivity and safety [6]; or to promote workers' good mental health [7]. In particular, the accurate recognition and segmentation of the actions, including the timing of when the actions commence and conclude, is essential for the cobot to understand and interpret the intended actions of the human collaborator, to synchronize its actions, respond in real-time, and ensure smooth cooperation with the human collaborator [8] [9].

Recently, the research has notably focused on using multimodal data, which can contribute to developing more sophisticated and adaptive action recognition systems. In particular, the information derived from skeletal joints enables researchers to capture temporal variations in body movements. It offers flexibility in focusing on the entire body or specific body parts, allowing for a comprehensive representation of the action recognition and bypassing eventual privacy concerns [10] [11].

To the best of the authors' knowledge, few vision-based datasets exist on human-cobot cooperation for object assembly in industrial manufacturing. For this reason, in the last few years, our research has been focused on the task of generating real datasets for practical applications of action recognition in the manufacturing context. The datasets for Human Action Multi-Modal Monitoring in Manufacturing (HA4M) and the Human-cobot collaboration for Action Recognition in Manufacturing Assembly (HARMA), consist of multimodal information acquired during the assembly of an Epicyclic Gear Train (EGT), depicted in Figure 1, without and with the collaboration of a cobot, respectively.

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
laura.romeo@stiima.cnr.it (L. Romeo);
annaclaudia.bono@stiima.cnr.it (A. Bono);
grazia.cicirelli@stiima.cnr.it (G. Cicirelli);
tiziana.dorazio@stiima.cnr.it (T. D'Orazio)
0000-0001-8138-893X (L. Romeo)

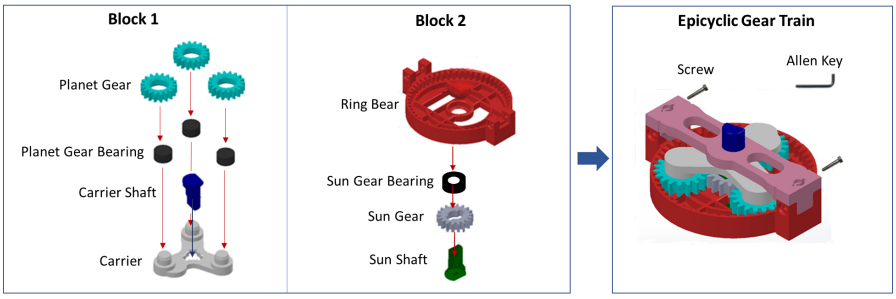

Figure 1: Components involved in the assembly of the Epicyclic Gear Train. The CAD model of the components is publicly available at [12].

The HA4M dataset was recorded using one single depth camera, while the HARMA dataset was recorded using two depth cameras. The Microsoft® Azure Kinects have been selected as depth cameras in both cases.

The two proposed datasets present various main contributions compared to the existing ones [13, 14] in the context of object assembly in industrial manufacturing:

- The datasets provide untrimmed sequences of several types of data: RGB frames, Depth maps, RGB-to-depth-Aligned (RGB-A) frames, and Skeleton data. The availability of a variety of multi-modal data represents an added value for the scientific community to test different machine learning approaches in action segmentation as well as action recognition tasks, by using one or more data modalities.
- The datasets present a variety in action execution due to the different order followed by the subjects to perform the actions and the interchangeable use of both hands.
- The actions have a high granularity as the components to be assembled and the actions themselves appear visually similar. As a result, recognizing different actions is very challenging and requires a high level of context understanding and object-tracking skills.
- Both datasets provide a good base for developing, validating, and testing techniques and methodologies for the recognition and segmentation of assembly actions.

Preliminary experiments have been conducted to test state-of-the-art temporal action segmentation methods, the ASFormer [15] and MS-TCN++ [16], on RGB and skeletal data achieving considerable accuracy rates in action segmentation.

The remainder of this paper is organized as follows: Section 2 presents the datasets and describes the assembly task, reporting details on the acquisition setup, study participants, and data annotation. Section 3 reports some experimental results on action segmentation. Finally, Section 4 delineates conclusive remarks.

2. Datasets description

The task involves the assembly of an Epicyclic Gear Train (EGT) (see Figure 1), which involves three phases: the assembly of Block 1, the assembly of Block 2, and then the completion of the EGT that makes up both blocks. The HA4M dataset contains videos of different operators that assemble the complete EGT. The HARMA dataset, instead, contains videos of different operators that assemble the EGT in collaboration with a cobot. All the subjects participated voluntarily in the experiments. They were asked to execute the task several times as preferred (e.g. with both hands), independently of their dominant hand. Furthermore, the subjects performed the task at their comfortable self-selected speed so that high time variance could be noticed among the different subjects. The subsequent sections give more details on both datasets.

2.1. HA4M dataset

The HA4M dataset contains 217 videos of the assembly task performed by 41 subjects. The acquisition setup is composed of a Microsoft Azure Kinect® camera placed on a tripod in front of the operator as pictured in Fig. 2.

The camera is at a height of $1.54\,m$ above the floor, at a horizontal distance of $1.78\,m$ from the far border of the table, and is tilted down to an angle of $17°$. As shown in Figure 2, the individual components to be assembled are spread on the table in front of the operator and are placed according to the order of assembly. The operator can pick up one component at a time to perform the assembly task standing in front of the table. The experiments took place in two laboratories: one in Italy and one in Spain. Two typical RGB frames captured by the camera in both laboratories are shown in Figure 3. The Figure also depicts the two supports fixed on the table to facilitate the assembly of Block 1 and Block 2.

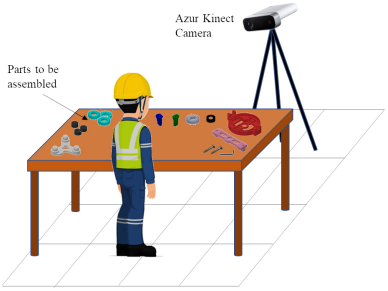

Figure 2: Sketch of the acquisition setup of the HA4M dataset: a Microsoft® Azure Kinect is placed in front of the operator and the table where the components are spread over.

(a) (b)

Figure 3: Typical video frames acquired by the RGB-D camera in the (a) Italian and (b) Spanish Laboratories.

Table 1
List of Block 1, Block 2, and EGT components, respectively.

EGT Components		
	Quantity	**Description**
Block 1	3	Planet Gear
	3	Planet Gear Bearing
	1	Carrier Shaft
	1	Carrier
Block 2	1	Ring Bear
	1	Sun Gear Bearing
	1	Sun Gear
	1	Sun Shaft
EGT	1	Block 1
	1	Block 2
	1	Cover

Tables 1 and 2 list the components and the actions necessary for assembling Block 1, Block 2, and the whole EGT, respectively. Notice that the final action (ID=12) involves additional tools, such as two screws and an Allen key to secure the EGT.

As listed in Table 2, the total number of actions is 12, divided as follows: four actions for building Block 1, four for building Block 2, and four for assembling the two blocks and completing the EGT. Some actions are performed more times as there are more components of the same type to be assembled: actions 2 and 3 are executed three times, while action 11 is repeated two times. Finally, a "don't care" action (ID=0) has been added to manage pauses between action transitions or unexpected events such as the loss of a component during the assembly.

Table 2
List of actions to build Block 1, Block 2, and EGT in the HA4M dataset.

Actions		
	ID	**Description**
	0	"don't care" action
Block 1	1	Pick up/Place Carrier over Support 1
	2	Pick up/Place Gear Bearings (×3)
	3	Pick up/Place Planet Gears (×3)
	4	Pick up/Place Carrier Shaft
Block 2	5	Pick up/Place Sun Shaft over Support 2
	6	Pick up/Place Sun Gear
	7	Pick up/Place Sun Gear Bearing
	8	Pick up/Place Ring Bear
EGT	9	Pick up Block 2 and place it on Block 1
	10	Pick up/Place Cover
	11	Pick up/Place Screw (×2)
	12	Pick up Allen Key, Turn both screws, Return Allen Key and the EGT

2.2. HARMA dataset

The HARMA dataset comprises 160 videos (80 videos per camera) capturing the assembly task performed by 27 subjects in collaboration with a cobot (Fanuc CRX10ia/L robotic arm). Each subject performed the task multiple times, resulting in 240 task executions in the dataset.

The acquisition setup is pictured in Fig. 4. The two Microsoft® Azure Kinect cameras are placed on a tripod in Frontal and Lateral positions to the Operator Workplace. The Frontal Camera is at a height of $1.72\,m$ above the floor and down tilted by an angle of $6\,degrees$, while the Lateral Camera is at a height of $2.07\,m$ and $19\,degrees$ down tilted. Two typical RGB frames captured by both cameras are shown in Fig. 5. As shown in Fig. 5, the EGT components are spread over the Operator Workplace, so the operator can pick up one component at a time to perform the assembly task in seven pick-and-place actions [14]. The operator assembles Block 1, whereas the cobot assembles Block 2. The assembly of Block 2 done by the cobot is not considered in the HARMA dataset, as our goal is to recognize the actions performed by the operator to trigger the cobot when it

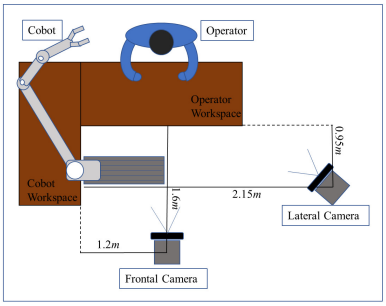

Figure 4: Sketch of the acquisition setup of the HARMA dataset: two Microsoft® Azure Kinect cameras are placed in a Frontal and Lateral position to the operator's workplace.

(a) (b)

Figure 5: Sample frames captured by the (a) Frontal and (b) Lateral camera, respectively, during the assembly task.

has to approach the operator to perform the collaboration action. So, the HARMA dataset comprises videos of only the assembly task performed by the subjects, including the collaborative action needed to join Block 1 and Block 2 (action 5 in Tab. 3). Table 3 lists the seven actions included in the HARMA dataset. As can be noticed in Table 3, unlike the HA4M dataset, the Cover is secured with two hooks (see Figure 6).

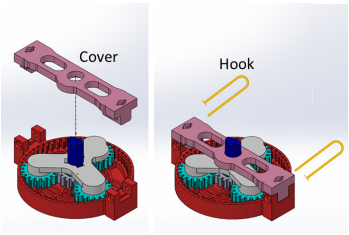

Figure 6: Completion of the EGT by placing the Cover and the two Hooks as included in Action 7 of Table 3.

3. Experiments

This section presents preliminary experiments and results on temporal action segmentation by applying state-of-the-art deep learning methods to HA4M and HARMA datasets. Both datasets were split into non-overlapping training and testing sets by considering the 70% of videos for training and the remaining 30% for testing, ensuring that videos of the same operator do not appear in both training and testing sets.

Table 3
List of the actions carried out by the operator for the construction of the EGT in the HARMA dataset.

Actions		
	ID	**Description**
	0	"don't care" action
Block 1	1	Pick up/Place Carrier over the Support
	2	Pick up/Place Planet Gear Bearing (×3)
	3	Pick up/Place Planet Gear (×3)
	4	Pick up/Place Carrier Shaft
EGT	5	Pick up Block 1 and join it with Block 2 held by the cobot
	6	Pick up/Place the Cover
	7	Pick up/Place the 2 Hooks, then leave the EGT on the table

ASFormer [15] and MS-TCN++ [16] models have been applied to test action segmentation performance. The ASFormer (risp. the MSTCN++) models were fed using RGB and Skeletal data extracted from both datasets, performing the training over 120 (risp. 100) epochs, collecting losses for each iteration. The best model is chosen as the one with the lower loss within the total number of iterations and is used in the test phase.

Tab. 4 lists the performance rates in terms of Accuracy, Edit Score, and F1-score. Accuracy is a frame-wise metric that measures the proportion of correctly classified frames in the entire video sequence without capturing the temporal dependencies between action segments. The Edit Score, instead, measures how well the model predicts the ordering of action segmentation without requiring exact frame-level alignment. Finally, F1-score with a threshold τ, often denoted as F1@τ, accounts for the degree of overlap between the Intersection over Union (IoU) of each predicted segment and ground truth segments [17]. In the experiments, the threshold τ has been set to 60%, 70% and 80%. Focusing on these metrics, it can be noticed that all the considered models succeeded in correctly segmenting the actions for the assembly task. In particular, the Accuracy rates reached high values (over 91%) in both cases of using RGB or skeletal features.

For completeness, Figure 7 shows a qualitative representation of action segmentation obtained by applying MS-TCN++on and ASFormer models to one video from the HA4M and one from the HARMA dataset. These

Table 4
Performance rates on action segmentation obtained by applying ASFormer and MS-TCN++ architectures, using RGB and Skeletal data grabbed from HA4M and HARMA datasets.

TAS Model	*Dataset*	*Features*	*Acc.*	*Edit*	*F1 @ {60, 70, 80}*		
ASFormer [15]	HA4M	RGB	91.79%	95.10%	87.81%	80.82%	70.27%
		Skeleton	92.43%	93.01%	86.71%	79.28%	69.42%
	HARMA	RGB	94.2%	93.6%	92.0%	88.7%	83.4%
		Skeleton	94.51%	95.08%	91.03%	87.97%	78.24%
MS-TCN++ [16]	HA4M	RGB	93.53%	93.85%	91.12%	86.01%	76.22%
		Skeleton	94.92%	95.9%	92.57%	88.57%	81.85%
	HARMA	RGB	92.13%	86.23%	78.18%	74.54%	66.00%
		Skeleton	94.45%	93.89%	90.24%	87.80%	81.80%

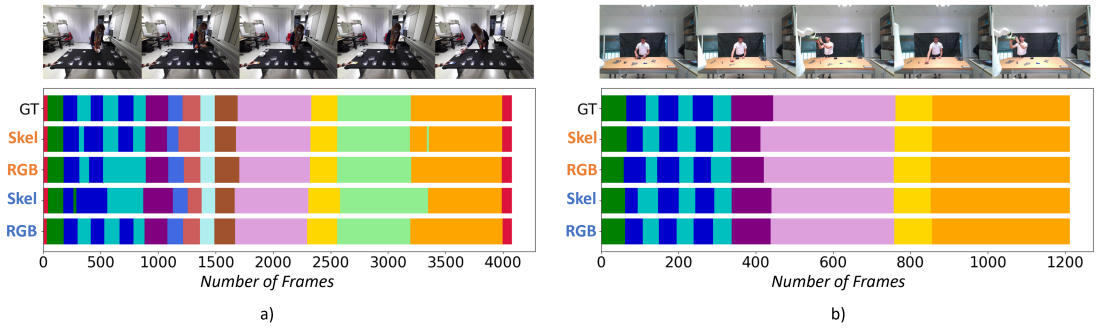

Figure 7: Action segmentation results over a video from the HA4M (a) and a video from the HARMA (b) dataset. GT, RGB, and Skel stand for Ground Truth, use of RGB features and use of Skeletal features, respectively. The labels in orange indicate the results obtained by the MS-TCN++ model, while the labels in blue remark the outcomes of the ASFormer architecture.

videos have been chosen to display challenging situations such as the case of Action2 (dark blue bars) and Action3 (light blue bars) that in the case of HA4M (Fig. 7(a)) are not always detected properly depending on the used features or applied model. On the contrary, Fig. 7(b) shows better segmentation results also for actions 2 and 3. Furthermore, in the HARMA dataset, the availability of two cameras allows us to compensate for the lack of data when one camera fails to provide skeletal data due to occlusion or out of range [18].

4. Conclusions

The present paper depicted an examination of two industrial datasets, namely the Human Action Multi-Modal Monitoring in Manufacturing (HA4M), and the Human-cobot collaboration for Action Recognition in Manufacturing (HARMA). Both datasets address the high demand for human action recognition and segmentation within industrial manufacturing contexts, particularly regarding scenarios involving Human-Robot collaboration and interaction. The multimodal features within the datasets encompass a variety of actions and interactions in industrial assembly tasks, allowing this work to lay the foundation for the development and enhancement of intelligent systems aiming at the understanding and assisting human operators in manufacturing production lines.

To properly evaluate HA4M and HARMA, state-of-the-art temporal action segmentation models were considered, namely ASFormer and MS-TCN++, which demonstrated notable success in exploiting the data provided by the datasets. The comparison between the RGB and Skeletal features underlines the potential of a multimodal approach to balance the computational efficiency with the precision required for the recognition and segmentation of complex tasks.

The conducted experiments prove that, overall, both RGB and Skeletal features performed properly. RGB data provides rich visual information about the scene but typically requires higher storage space and computational

complexity compared to skeleton-based data representation. On the other hand, by using skeleton data is possible to abstract away detailed appearance information and focus solely on the spatial configuration of body joints and movements. Therefore, it's essential to carefully find a good trade-off and select the data modality that best aligns with the goals and constraints of the working context.

The presented datasets are benchmarks for further studies in novel models and algorithms that can improve the accuracy and reliability of action recognition and segmentation systems in industrial settings. HA4M and HARMA offer a valuable resource for the research community, allowing ongoing innovation and development of human-robot collaboration systems in complex, real-world scenarios.

Acknowledgments

This research has been partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI", funded by the European Commission under the NextGeneration EU program.

References

[1] A. Keshvarparast, D. Battini, O. Battaia, A. Pirayesh, Collaborative robots in manufacturing and assembly systems: literature review and future research agenda, Journal of Intelligent Manufacturing (2023).

[2] L. Wang, R. Gao, J. Vancza, J. Krüger, X. Wang, S. Makris, Symbiotic human-robot collaborative assembly, CIRP Annals - Manufacturing Technology 68 (2019) 701–726.

[3] W. Tao, M. Al-Amin, H. Chen, M. C. Leu, Z. Yin, R. Qin, Real-Time Assembly Operation Recognition with Fog Computing and Transfer Learning for Human-Centered Intelligent Manufacturing, Procedia Manufacturing 48 (2020) 926–931.

[4] J. Patalas-Maliszewska, D. Halikowski, R. Damaševičius, An Automated Recognition of Work Activity in Industrial Manufacturing Using Convolutional Neural Networks, Electronics 10 (2021) 1–17.

[5] M. A. Zamora-Hernandez, J. A. Castro-Vergas, J. Azorin-Lopez, J. Garcia-Rodriguez, Deep learning-based visual control assistant for assembly in industry 4.0, Computers in Industry 131 (2021) 1–15.

[6] T. Kobayashi, Y. Aoki, S. Shimizu, K. Kusano, S. Okumura, Fine-grained Action Recognition in Assembly Work Scenes by Drawing Attention to the Hands, in: 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2019, p. 440–446. doi:10.1109/SITIS.2019.00077.

[7] M. L. Nicora, E. André, D. Berkmans, C. Carissoli, T. D'Orazio, et al., A human-driven control architecture for promoting good mental health in collaborative robot scenarios, in: 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), 2021, pp. 285–291.

[8] G. Cicirelli, C. Attolico, C. Guaragnella, T. D'Orazio, A kinect-based gesture recognition approach for a natural human robot interface, International Journal of Advanced Robotic Systems 12 (2015).

[9] M. V. Maselli and R. Marani and G. Cicirelli and T. D'Orazio, Continuous Action Recognition in Manufacturing Contexts by Deep Graph Convolutional Networks, volume 825, Springer, 2024.

[10] L. Romeo, R. Marani, A. Perri, T. D'Orazio, Microsoft Azure Kinect Calibration for Three-Dimensional Dense Point Clouds and Reliable Skeletons, Sensors 22 (2022) 4986.

[11] C. Brambilla, R. Marani, L. Romeo, M. L. Nicora, F. A. Storm, G. Reni, M. Malosio, T. D'Orazio, A. Scano, Azure kinect performance evaluation for human motion and upper limb biomechanical analysis, Heliyon 9 (2023).

[12] D. F. Redaelli, F. A. Storm, G. Fioretta, MindBot Planetary Gearbox, 2021. URL: https://zenodo.org/record/5675810#.YZZJXrVKjcs. doi:10.5281/zenodo.5675810.

[13] G. Cicirelli, R. Marani, L. Romeo, M. G. Dominguez, J. Heras, A. G. Perri, T. D'Orazio, The HA4M dataset: Multi-Modal Monitoring of an assembly task for Human Action recognition in Manufacturing, Scientific Data 9 (2022).

[14] L. Romeo, R. Marani, G. Cicirelli and T. D'Orazio, A Dataset on Human-Cobot Collaboration for Action Recognition in Manufacturing Assembly, 2024. Submitted to CoDiT2024.

[15] F. Yi, H. Wen, T. Jiang, ASFormer: Transformer for Action Segmentation, in: The British Machine Vision Conference (BMVC), 2021.

[16] S.-J. Li, Y. AbuFarha, Y. Liu, M.-M. Cheng, J. Gall, MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 6647–6658.

[17] G. Ding, F. Sener, A. Yao, Temporal Action Segmentation: An analysis of modern techniques, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[18] L. Romeo, G. Cicirelli and T. D'Orazio, Multi-view skeleton analysis for human action recognition and segmentation tasks, 2024. Submitted to CASE2024.