Contents lists available at SciVerse ScienceDirect

CIRP Annals Manufacturing Technology

Journal homepage: www.elsevier.com/locate/cirp

Intelligent assembly operations monitoring with the ability to detect non-value-added activities as out-of-distribution (OOD) instances.

Vignesh Selvaraj [a], Md Al-Amin [b], Wenjin Tao [b], Sangkee Min (2)[a]

[a] *Mechanical Engineering, University of Wisconsin Madison, 1415 Engineering Drive, Madison, 53706, Wisconsin, USA*
[b] *Foxconn iAI, 120001 Braun Road, Mt Pleasant, 53177, Wisconsin, USA*

Recognition and localization of actions in manufacturing assembly operations enables improvements in productivity and product quality by identifying the bottlenecks and assembly errors. In our previous work, we developed an approach that can recognize and localize the assembly standard operating procedures (SOP) steps in real-time using vision cameras. In this work, we augment the previous study with the ability to detect objects corresponding to the step being performed. Additionally, identifying non-value-added (NVA) activities in an assembly operation is challenging, hence, in this study we propose an approach of detecting NVA activities by considering the out-of-distribution for deep learning models.

Monitoring, Assembly, Smart manufacturing

1. Introduction

A typical manufacturing assembly station consists of a series of Standard Operating Procedures (SOP) steps performed on every product that passes through it. In industries, up to 40% of the cost and 70% of the production time falls under assembly operations, either in intermediate assembly operations or in final finished product assemblies [1]. Hence, the quality of the assembly operations impacts the lead time and end-product quality by a huge margin.

Traditionally, assembly operation monitoring was performed manually, where a representative of assembly cycle time and step time were determined by observing a sample of assembly cycles. The manual approach presents a few challenges, firstly, the estimate could be biased as it is generally conducted over a smaller sample, secondly, the approach is time consuming and cannot be performed continuously, finally, errors in an assembly cycle cannot be detected in real-time. To overcome the challenges presented by the traditional approaches, several studies have been conducted that aim to recognize the activity in manufacturing assembly operations using hand-worn sensors [2-3]. These approaches use a combination of sensors to detect the actions performed by a human operator in an assembly workstation. These approaches perform well in detecting the actions but are unable to localize the detected actions. Additionally, in many manufacturing facilities, safety, and privacy concerns prevents the operators from wearing the sensors, and the authority to ensure that the sensors were operating effectively was up to the operators, which was not ideal. An approach to identifying human actions in an assembly by image processing has also been studied [4], but it focuses on assessing the completion of a process rather than determining step time and cycle time. A deep learning approach to recognizing human actions has also been studied [5]. Though this study does not directly relate to assembly monitoring, it explores the potential of using deep learning models in human action recognition.

In our work, we aim to develop a robust system that can reliably monitor a worker-centric assembly workstation in real time using vision cameras. The first phase of work involves determining the step time and cycle time and detecting anomalies for the assembly operations. The second phase of work involves the integration of object detection to guide the assembly operators. The third phase of work involves identifying the non-value-added (NVA) activities in an assembly operation. Identifying NVA activities is particularly challenging as NVA activities cannot be bounded by a class label for the deep learning models.

In this paper, we start by introducing our preliminary work in section 2. This is followed by the design of a human-centric assembly guidance system in section 3, where we discuss the integration of the object detection module to guide the assembly operators in real time. In section 4, we discuss an energy-based approach to identify NVA activities in an assembly workstation, which was a key concern when developing an autonomous assembly monitoring system. Finally, in section 5, we conclude our work and briefly describe future work. The scientific significance of this work was to develop a full-fledged assembly monitoring and guidance system, which can detect NVA activities without explicit training of the deep learning models.

2. State Machine Integrated Recognition and Localization (SMIRL)

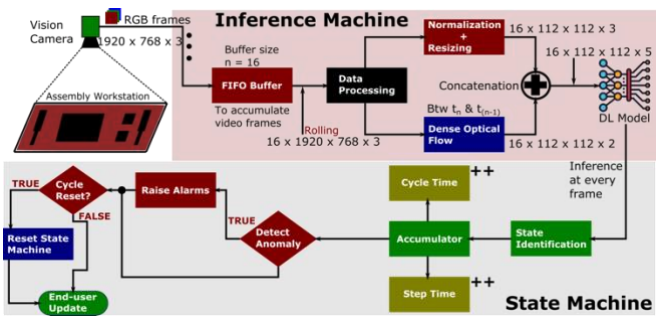

Figure 1. Inference Module of SMIRL

To continuously monitor the assembly operations, it is required to determine step time and cycle time for each assembly cycle. Additionally, it is also important to identify assembly operation anomalies like sequence breaks and missed steps. We developed

an approach called the State Machine Integrated Recognition and Localization (SMIRL) that can continuously monitor assembly workstations using vision cameras and determine the step times and cycle time of an assembly cycle in real time. The SMIRL comprises of an inference machine and a state machine, which synchronously work together to robustly detect and localize the actions, as can be seen in Fig. 1. The inference machine contains a deep learning model trained to detect the assembly step from video stream, whereas the state machine computes the time incurred at every assembly step and identifies anomalies.

3. Human-centric assembly guidance system

The assembly guidance system intends to guide the assembly operators in performing their tasks reliably, either during the training phase or in continuous operation. It also provides the means to ensure that the 5S practices (sort, set in order, shine, standardize, and sustain) are followed in the workstation without fail, to create an organized and productive workspace.

3.1. Evaluation of SMIRL

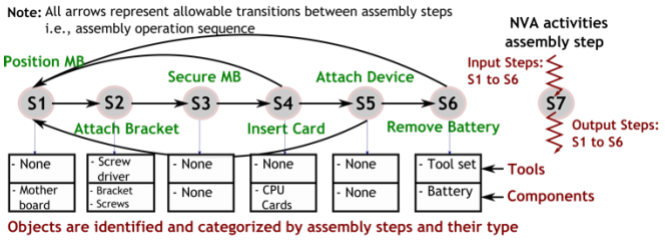

Figure 4. Relationship between assembly steps and its associated objects

In our previous study, to validate SMIRL, it was tested against an assembly operation that involves assembling a server motherboard onto a showcase for display. The assembly consists of 6 steps with an additional step added to collectively represent the NVA activities in the assembly. The relationship between the assembly steps and the sequence of operation can be seen in Fig.2. The metrics used to evaluate SMIRL were Normalize Mean Absolute Error (NMAE) and Intersection over Union (IoU). NMAE was computed between the human-inferred time and SMIRL-inferred time for each step in an assembly cycle. The IoU enables localizing the detect actions, by determining the overlap between the human-inferred and SMIRL-inferred assembly step time blocks. The results of the evaluation can be seen in Table 1.

Table 1 Evaluation of SMIRL

Assembly SOP steps	*NMAE*	*IoU*
Position Motherboard	0.07	0.8427
Attach Bracket	0.12	0.8959
Secure Motherboard	0.17	0.9055
Insert Card	0.07	0.9146
Attach Device	0.1	0.8336
Remove Battery	0.1	0.8596
Miscellaneous (NVA)	0.15	

3.2. Integration of object detection with SMIRL

To guide the assembly operators in performing their tasks, the objects: tools, and parts, corresponding to each assembly step need to be identified in real time. During the initialization of SMIRL, a state dependency matrix was provided that encapsulates the relationship between the assembly steps in an assembly operation.

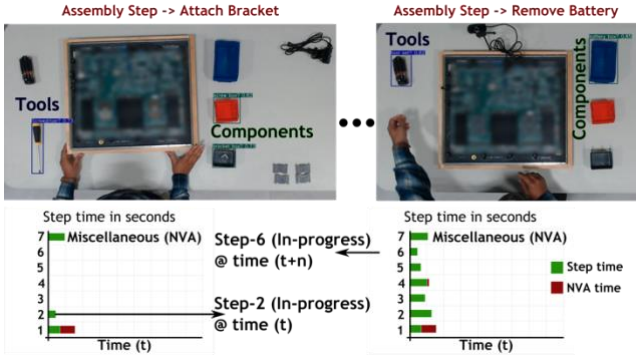

Figure 2. Integration of object detection and assembly action-localization

To detect objects, in addition to the state relationships, information on the tools and parts corresponding to each step was also provided. The state dependency matrix along with the objects associated with each assembly step can be seen in Fig. 2.

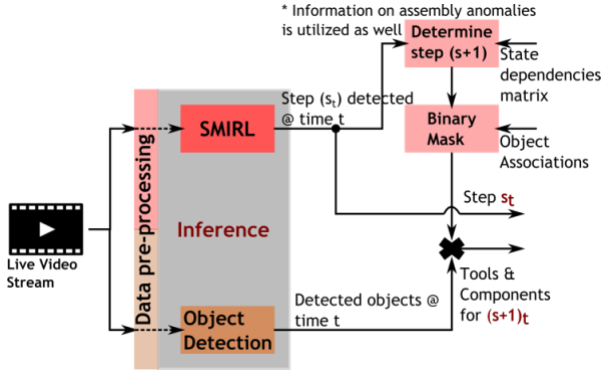

Figure 3. Inference process for the assembly guidance system

To realize a functioning guidance system, an object detection module was integrated with SMIRL. The model used to detect the objects was YOLOv7 [6], which was transfer learned by fine-tuning the layers to identify the objects corresponding to the assembly operation under study. During inference, the frames from the video stream were processed by both SMIRL and the object detection module in parallel. The frames input to the models were preprocessed. For the case of SMIRL, the preprocessing involves dense optical flow computation, normalization, and resizing the frames to the dimension 112×112. The optical flow augments motion detection from video frames, and normalization improves the deep learning model training process. The SMIRL identifies the assembly step s_t at time t. YOLOv7 identifies the objects present in the workstation. To guide the assembly operators, the objects corresponding to the next step, $(s+1)_t$, need to be identified. The inference from SMIRL and the state dependency matrix were used to identify the prospective next assembly step(s). Depending on the assembly step, $(s+1)_t$, a binary mask that highlights the tools and components was then generated. The binary mask selectively removes and/or highlights the objects corresponding to the state of the assembly cycle. The objects corresponding to assembly tools were highlighted using blue bounding boxes, and the components that go into the part were highlighted using green bounding boxes, see Fig. 3. The described inference process can be seen in Fig. 4. The developed assembly guidance system can infer at a rate of 55 FPS using a single RTX-6000 GPU. In addition to guiding the assembly operators in performing their tasks, it can also ensure that the 5S standards, particularly "Set in Order", was maintained at the workstation. By selectively identifying the tools required to perform an assembly step, it was ensured that they were "Set in Order" once the tasks were complete. A snapshot of the inference process showing the assembly step inferred by SMIRL and the identified objects can be seen in Fig. 3.

3.3. Assembly guidance system

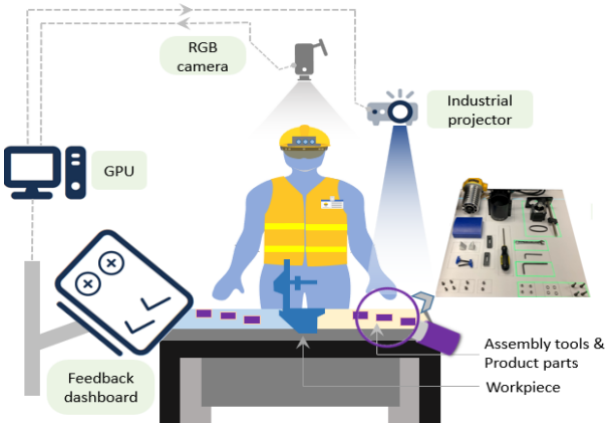

Figure 5. AI integrated assembly guidance system

The framework of AI integrated assembly guidance system is shown in Fig. 5. The system is comprised of four modules: 1) SMIRL, 2) Object detection system, 3) Projection system, to overlay instructions directly on the workbench, and 4) Feedback module to rectify detected anomalies. The camera overlooking the assembly workstation streams the assembly operation. The captured frames are sent to the SMIRL and object detection model for inference. Using the results, the tools and parts on the workstation are highlighted using a projector. The camera and projector view are calibrated to ensure the location of the detected bounding box match projected one. A set of performance measures along with the feedback to rectify the assembly mistakes (sequence break, missed step) are shown using a monitor attached to the workbench.

4. Identifying NVA activities through OOD detection

The key challenge with SMIRL was the inability to identify NVA activities in an assembly operation without explicitly training the deep learning models on them. The NVA activities correspond to all the activities performed in an assembly workstation that does not correspond to any of the assembly SOP steps. It is not possible to train the deep learning models on NVA activities in an assembly operation, as they could be n-folds, and identifying all scenarios is not possible. Hence in this study, we developed an approach where the NVA activities were identified as an out-of-distribution (OOD) instance. The approach was then evaluated for their ability to accurately identify the NVA activities at each assembly step.

The data used in this study were collected by recording several cycles of the assembly of a motherboard to its showcase. During the assembly operation, NVA activities were synthetically generated by letting the assembly operator to be idle or performing tasks other than assembly SOP steps. The data were recorded across two different operators performing the same assembly step. Throughout the rest of the paper, the terminologies NVA activities and OOD were used synonymously.

4.1. Energy-based model for NVA activities detection

The basis of this approach was to use energy-based modeling to map each point of the input space to a single non-probabilistic scalar called the energy. An energy bounded learning objective was used to explicitly create an energy gap between the in-distribution and out-of-distribution data [7]. The objective function used in the model training process can be seen in Eq. (1). The overall training objective is the combination of cross entropy loss and the regularization loss in energy, Eq. (2).

$$min_\theta \; E_{(x,y)\sim\mathcal{D}_{in}^{train}}\left[-log\mathcal{F}_y(x)\right] + \lambda \cdot \mathcal{L}_{energy} \quad (1)$$

$$\mathcal{L}_{energy} = E_{(x_{in},y)\sim\mathcal{D}_{in}^{train}}\left(max(0, E(x_{in}) - m_{in})\right)^2 + E_{(x_{nva})\sim\mathcal{D}_{nva}^{train}}\left(max(0,\; m_{nva} - E(x_{nva}))\right)^2 \quad (2)$$

where $\mathcal{D}_{in}^{train}$ corresponds to the in-distribution training data, corresponding to the assembly SOP steps, and $\mathcal{D}_{nva}^{train}$ was the auxiliary data corresponding to the NVA activities. For the first term in $\mathcal{L}_{energy}$, the model penalizes the samples corresponding to the assembly steps that produces energy higher than m_{in}. Similarly, for the second term, the model penalizes the samples corresponding to NVA activities that produces energy lower than m_{nva}. The m_{in} and m_{nva} are hyperparameters and need to be determined experimentally.

After pre-processing, the data for the model training process had the dimension $16 \times 112 \times 112 \times 5$, where the first dimension was temporal length, second and third dimensions referred to the width and height of frames, and the fourth dimension was the concatenation of RGB frames and Optical Flow information. The terms, $E(x_{in})$ and $E(x_{nva})$ represent the expectation along the temporal length dimension. The dimension to compute the expectation was experimentally evaluated, and temporal length dimension was found to be best performing in terms of creating the energy gap between assembly steps and NVA activities.

During the evaluation, for every input, the energy was computed as seen in Eq. (3). Then using the predetermined threshold δ, the inputs were classified as assembly SOP steps or NVA activities.

$$E(\mathrm{x}; f) = -T \cdot \log \sum_{\mathrm{i}}^{\mathrm{K}} \mathrm{e}^{\mathrm{f_i(x)/T}} \quad (3)$$

In Eq. (3), T is the temperature scaling factor, $f_i(x)$ are the logits from the last dense layer of the deep learning model, K is the number of logits in the penultimate layer of the neural network classifier. After computing the energy, the input data were classified between VA (value-added), class-0, and NVA (non-value-added), class-1, using an experimentally determined threshold as shown in Eq. (4).

$$g(\mathrm{x}; \tau, f) = \begin{cases} 0 & if \;\; E(\mathrm{x}; f) \;\leq\; \tau \\ 1 & if \;\; E(\mathrm{x}; f) \;>\; \tau \end{cases} \quad (4)$$

The notion of using a regularizer, such as $\mathcal{L}_{energy}$, was to push the optimization objective for deep learning models, Eq. (1), to create a gap in the dimensionless energy, $E(\mathrm{x}; f)$, between the in-distribution and NVA data instances.

4.2. Evaluation of NVA activities detection

The deep learning model in the inference machine of SMIRL was trained using 13 assembly cycles, with an average cycle time of 107.5 ± 7.9s. Each step of the assembly operation was manually labelled, including the NVA activities in each of the cycles in the training data. The training data constitutes 6 assembly steps, with one more added to represent all NVA activities. Based on the annotation information, the data were categorized into $\mathcal{D}_{in}^{train}$ and $\mathcal{D}_{nva}^{train}$. The hyperparameters used in the energy score computation, T, m_{in}, and m_{nva}, were identified using grid searching approach, and were determined to be 500, -5, and -27, respectively. The trained model was then evaluated on 206 unseen assembly cycles. The distribution of the energy computed across all these cycles for all assembly SOP steps were then plotted in Fig. 6. From the distribution, the threshold for the energy between assembly SOP steps and NVA activities was determined. Using the determined threshold, the input to the model was classified between assembly SOP steps, all 6 of them, or NVA activities. The F1-score for this binary classification was determined to be 0.9744, with precision and recall being 0.9799 and 0.9689, respectively.

4.3. Evaluation of assembly step time computation

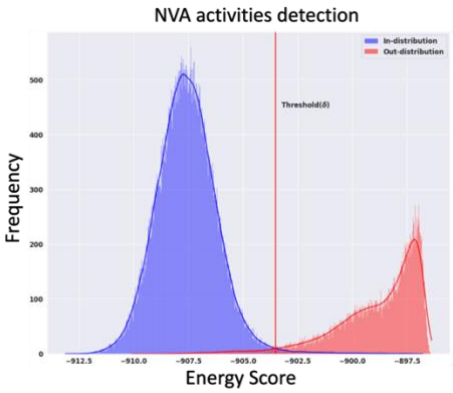

Figure 6. Distribution of energy scores for SOP steps and NVA activities

In this section, the NVA activities detection was evaluated by integrating OOD detection approach with SMIRL. The assembly step time was determined for the cases of with and without OOD detection and was compared with the human inferred step time using NMAE, Eq. (5). The results can be seen in Table 2. Without the energy-based approach for identifying NVA activities, all the NVA activities in each assembly cycle were detected to be step-1 (Position Motherboard), leading to its step time being 3-folds the human annotated time.

$$NMAE = \frac{abs\left(Inferred_{st_i} - Actual_{st_i}\right)}{\mu_{st_i}} \quad (5)$$

Table 2 NMAE after NVA activities identification

Assembly SOP steps	*Without NVA detection*	*With NVA detection*
Position Motherboard	3.05	0.15
Attach Bracket	0.12	0.12
Secure Motherboard	0.2	0.21
Insert Card	0.15	0.15
Attach Device	0.07	0.07
Remove Battery	0.36	0.14
Miscellaneous (NVA)	-	0.18

On contrast to approach in Table 1, where the model was exclusively trained on NVA activities, the current approach requires far less data with only a slight loss in performance. Additionally, the NVA activities across different assembly workstation in an assembly line can be collated to represent $\mathcal{D}_{nva}^{train}$ distribution of the assembly line.

The proportion of NVA activities used as $\mathcal{D}_{nva}^{train}$ can impact its detection ability. Hence, the OOD detection ability was evaluated, by computing the F1-Score and the Area Under Receiver Operating Characteristic (AUROC) curve, for different proportions of NVA activities, as seen in Table 3. The ROC curve is a plot between true positive rate and false positive rate, and AUROC summarizes the curve into a single number between 0 and 1. The proportion here corresponds to the ratio of total time spent in NVA activity to total time spent working on the assembly SOP steps, across all 13 cycles used in the model training.

Finally, the sensitivity of the developed approach to changes in lighting conditions, anthropometric variations of the assembly operators, and speed of the assembly operation were studied. From our study, the following deductions were made: 1) The anthropometric variations associated with the operators and speed of motion had minimal impact. The reason can be attributed to the diversity in the training dataset, which enabled the model to form a resistance against anthropometric variations. Also, the higher camera FPS enabled in capturing rapid human actions reliably. 2) The most impactful factor was the lighting of the workstation; this is because the optical flow relies on the lighting intensity variation between frames. Hence, in this work, we assumed the lighting conditions of the workstations to be unchanged during their operation and monitoring.

Table 3 Proportion of NVA activities and its impact on OOD detection

Proportion of NVA activities	*F1-Score*	*AUROC*
3.2%	0.8321	0.8752
14.3%	0.8666	0.8965
24.5%	0.9373	0.9495
51.8%	0.9744	0.9804

5. Conclusion and future work

Detection and localization of actions corresponding to the assembly SOP steps helps in reducing costs and improving productivity. A real-time assembly guidance system can improve the operator's performance and enable effective training for new operators. The insights from this work are in three folds:

(1) A real-time assembly guidance system was developed that can identify the assembly step being performed and guide the operators by identifying the required tools and components.

(2) Identifying NVA activities is challenging. An approach was developed and studied to effectively identify NVA activities with an F1-Score of 0.9744 with no explicit DL model training.

(3) The developed approach was evaluated against the traditional approach of supervised learning of NVA activities. The performance of the system for different proportions of NVA activities was also evaluated.

As the future work, the NVA activities detection on pre-trained models will be studied, thereby eliminating the need for re-training. In a typical assembly line consisting of multiple assembly workstation, it is beneficial to transfer the knowledge on NVA activity detection across the assemblies, hence the correlation between the $\mathcal{D}_{nva}^{train}$ and the assembly operation performed will be studied.

Acknowledgement

This material is based on work supported by the National Research Foundation of Korea (Brain Pool Program 2022H1D3A2A01093491). The authors of this work would like to acknowledge Foxconn iAI, a division of Foxconn, for their support in providing the data required to perform this study.

References

[1] Aehnelt, M., Gutzeit, E., Urban, B., 2014, Using Activity Recognition for the Tracking of Assembly Processes: Challenges and Requirements, WOAR, 2014, 12–21.

[2] Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., Amirat, Y., 2015, Physical Human Activity Recognition Using Wearable Sensors, Sensors, 15, 31314–31338.

[3] Koskimaki, H., Huikari, V., Siirtola, P., Laurinen, P., Roning, J., 2009, Activity recognition using a wrist-worn inertial measurement unit: A case study for industrial assembly lines, 2009 17th Mediterranean Conference on Control and Automation, 2009, 401–405.

[4] Urgo, M., Tarabini, M., Tolio, T., 2019, A human modelling and monitoring approach to support the execution of manufacturing operations, *CIRP Annals*, 68, 5–8.

[5] Wang, P., Liu, H., Wang, L., Gao, R., X., 2018, Deep learning-based human motion recognition for predictive context-aware human-robot collaboration, *CIRP Annals*, 67, 17–20.

[6] Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M., 2022, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, arXiv preprint arXiv:2207.02696.

[7] Liu, W., Owens, J. D., Wang, X., Li, Y., 2020, Energy-based Out-of-distribution Detection, Advances in Neural Information Processing Systems, 33, 21464–21475.