Automated Progress Monitoring in Modular Construction Factories Using Computer Vision and Building Information Modeling

R. Panahi[1], J. Louis[2], A. Podder[3], S. Pless[4], C. Swanson[5], and M. Jafari[6]

[1,2] School of Civil and Construction Engineering, Oregon State University, Oregon, USA
[3,4] National Renewable Energy Laboratory, Colorado, USA
[5] Momentum Innovation Group, USA
[6] College of Built Environments, University of Washington, Washington, USA

E-mail: panahir@oregonstate.edu, joseph.louis@oregonstate.edu, ankur.podder@nrel.gov, shanti.pless@nrel.gov, colby@miginnovation.com, melikaj@uw.edu

Abstract –
Modular construction methods have recently gained interest due to the advantages offered in terms of safety, quality, and productivity for projects. In this method, a significant portion of the construction is performed off-site in factories where modular components are built in different workstations, assembled on the production line, and shipped to the site for installation. Due to the labor-intensive nature of tasks, cycle times in modular construction factories are highly variable, which commonly leads to major bottlenecks and delays in construction projects. To remedy this effect, recent methods rely on sensors such as RFID to monitor the production process, which is reportedly expensive, and intrusive to the work process. Recently, computer vision-based methods have been proposed to track the production process in modular construction factories. However, these methods overlook monitoring the assembly process on the production line. Therefore, this paper presents a method to monitor the assembly process by integrating computer vision-based methods with Building Information Modeling (BIM). The proposed method detects the modular units using object segmentation; superimposes the installation area with the corresponding 2D region using BIM, and identifies the installation of the components using image processing techniques. The proposed method has been validated using surveillance videos captured from a modular construction factory in the US. Successful implementation of the proposed method can lead to timely identification of delays during the assembly process and reduce delays in modular integrated construction projects.

Keywords –
Modular Construction; Computer Vision; Building Information Modeling

1 Introduction

Off-site and modular construction is increasingly being seen as a promising method of project delivery due to the advantages offered in terms of productivity, schedule, and cost [1]. Here, building components are produced in a controlled environment such as a prefabrication factory, and shipped to the site for installation, leaving relatively minimal work to be performed on-site [2]. It also enables incorporation of energy-efficient building strategies at scale to reduce first cost of installation for affordable housing. Despite these advantages and the recent advancements in the application of robotics inside modular construction factories [3], [4], the current state of these factories in the U.S. still highly relies on manual labor. In addition, variability in design [5] and the stochastic nature of the orders [6], all, lead to the high variability of cycle times and result in major bottlenecks. These bottlenecks reportedly can account for up to 15% of work time, reduce productivity, and cause delays in modular integrated construction projects [7]. Therefore, it is important to identify and mitigate these bottlenecks in order to prevent such delays from negatively affecting construction projects.

Monitoring and control systems inside factories can be used to identify bottlenecks inside the factories and enable optimal tactical and strategic responses to dynamic changes on the factory floor [8]. However, the current state of the monitoring practice inside modular

construction factories commonly relies on manual methods such as five-minute rating, and work sampling [9], which are prohibitively expensive and error-prone when implemented in a large scale [10]. Alternatively, sensors such as radio-frequency identification (RFID) [5], audio [11], and inertial measurement units (IMU) [12] were used to automatically collect the process data from the shopfloor in modular construction factories. However, these systems are expensive to maintain and can cause intrusion into the progress of work, which can impede their practical application [13].

Computer vision-based methods can overcome these challenges by remotely monitoring the progress of work. During the past decade, these methods have attracted much attention from the construction community at large [14]. However, despite these advancements, recent research unveils an array of challenges computer vision-based methods face when applied inside modular construction factories. Examples of such challenges are the fast-paced environment which causes inter-object occlusions [15], high variability of tasks which reduces the performance of activity recognition methods [16], and fundamental differences in processes compared to on-site tasks which hinder the vision-based knowledge transfer across on-site and off-site environment [17].

To address these challenges, computer vision-based methods were used to monitor the ergonomics of workers [18], their activities [15], and tackle technical challenges such as occlusion in tracking the workers [16]. To monitor the processes, Zheng et al. [19] proposed a framework that extracts the cycle time for the installation of modules after being delivered to the construction site. They fine-tuned a Mask R-CNN object segmentation algorithm on a total of 1100 synthetic and real images of finished prefabricated modules. In order to monitor the progress of work in panelized construction factory workstations, Martinez et al. [17] proposed a vision-based monitoring method that detects the crane and the workers in a single station and updates the parameters of a finite state machine to track the progress of work. More recently, Park et al. [20] [21] created a synthetic image dataset of modular units inside the factory and evaluated a CNN-based 3D reconstruction network from the collected 2D synthetic images.

Based on the conducted review of the related literature following gaps are identified and targeted in this study: (1) drawbacks of contact-based sensors: majority of the previously proposed monitoring methods inside modular construction factories rely on contact-based sensors. Despite the accuracy these methods provide, they are expensive to implement at large scale; they are susceptible to noise, and they are intrusive to the progress of work; (2) drawback of vision-based monitoring methods: previous research identifies the movement of equipment and uses this information as queues to monitor the progress of work. However, in many cases the movement of equipment is not related to the progress of work, such as the assembly process, and (3) limitation of detection vs. segmentation: previously proposed progress monitoring methods inside modular construction factories rely on detection methods, however, detection bounding boxes are not reliable for progress monitoring as they entail a large background area, unrelated to the object of interest, especially in oblique views of CCTV video footage.

This study attempts to propose a computer vision-based progress monitoring method to overcome the drawbacks of contact-based sensors, use object segmentation to detect the modular units at pixel level, and monitor the assembly progress of components on the modular unit on the production line.

The remainder of this paper is structured as follows. First, the proposed method is presented. This is followed by a brief explanation of a case study that demonstrates the applicability of the framework. The paper ends with conclusions and future work.

2 Methodology

The goal of the proposed method is to monitor the assembly progress in modular construction factories. Figure 1 shows the objectives to achieve this goal.

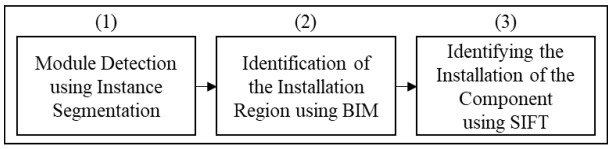

Figure 1. Workflow of the proposed method

As shown in Figure 1, the objectives of this study are: (1) module detection: to detect and segment the modular units on the shop floor; (2) identification of the installation region: to detect the installation region of interest (RoI) on the video, and (3) identifying the installation of the component: to classify the state of the components as installed or not-installed. These objectives are explained in further detail, in following sections.

2.1 Module Detection

Here, the modular units are detected and their boundary is segmented using an object detection and segmentation method. Doing so, Mask R-CNN instance segmentation is used to demarcate the modular units on the shop floor, at a pixel level. Mask R-CNN is a state-of-the-art model for instance segmentation, developed on top of Faster R-CNN. Faster R-CNN is a region-based convolutional neural networks [22], that returns

bounding boxes for each object and its class label with a confidence score. Figure 2 shows the architecture of the Mask R-CNN algorithm used on this study.

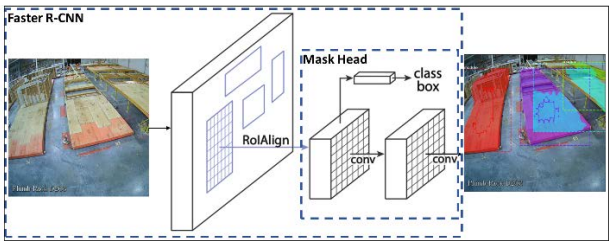

Figure 2. Architecture of the Mask R-CNN algorithm

As shown in figure 2, the algorithm first performs detection by drawing a bounding box around the object of interest. However, this bounding box commonly includes a large portion of the background area, since CCTV cameras are commonly installed in highly oblique orientation to cover a large portion of the factory. As a result, these bounding boxes cannot be efficiently used to extract precise reference points on the modular unit and use them to identify the installation regions. Therefore, the Mask Head in the segmentation algorithm is used to precisely annotate the boundaries of the modular unit. Finally, in order to identify the left corner of the module, the lowest point in each detected instance is identified as the reference point.

2.2 Identification of Installation Region

Here, the installation region of each component on the modular unit is annotated in the video using the BIM model. Figure 3 shows the pipeline that creates this association between the points in the virtual and real space using projective geometry, and equation 1 shows the projective transformation formula using the homography matrix H.

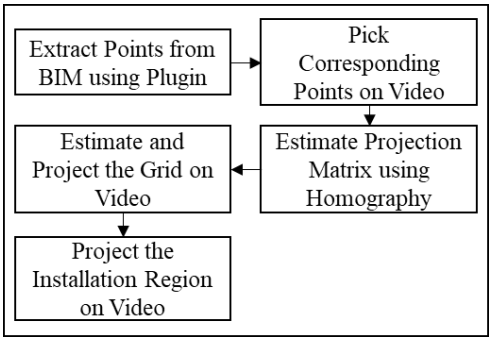

Figure 3. Region of Interest Identification

$$\begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (1)$$

In equation 1, (x_1', x_2', x_3') denotes the pixel coordinates of a point on the image, (x_1, x_2, x_3) denotes the pixel coordinates of a point on in the virtual space, and h_1 to h_9 denote the translation, rotation , and scale parameters. In this case, x_3', x_3 are set to zero as both sets of points are constrained to planes.

As shown in Figure 3, a plugin is designed in Revit software to extract four points located on the modular unit. Specifically, the designed plugin uses Revit API to loop through all the coordinates of a the "floor" type, and exports the four corners of the element to an excel sheet. Subsequently, these four corner points are manually identified on the video and the projection matrix is estimated using the homography formula shown in equation 1. It is important to note that this process needs to be done only once for each CCTV camera, since all the future in-coming modules are at the same height and therefore on the same plane. Therefore, the same homography matrix can be used for all stations visible in the same camera. However, the module detection step and identifying a reference point on the module has to be performed each time a module comes into the station since the exact location of the module inside the station is not fixed. Furthermore, the simple template-based matching expands the generalizability and practicality of this method to other factories by allowing the domain experts to easily annotate two templates for each station in the new factory. This approach also made this method robust to the changes in the appearance of the modular unit which occurs across the production line as the modular units are developed.

A grid parallel to the edges of the modular unit is projected on the plane of the modular units in the video, using the corner of the unit as the reference point. Finally, extracted points of the component, which needs to be installed, are projected on the generated grid. Equation 2 shows how the homographic matrix is used to calculate the pixel location of installation regions and what input BIM points are used.

$$\begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix}^{-1} \begin{bmatrix} X \\ Y \\ 0 \end{bmatrix} = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} \quad (2)$$

In Equation 2, the inverse of the H matrix, computed from the previous equation, is used and multiplied by (X, Y, 0), which denotes the coordinates of one corner of the region of interest from the BIM model. In the proposed method, the (X, Y, 0) coordinates are exported from the BIM model, where the designed API performs a search for elements with type "bath pod" and returns the corner

coordinates of the element. Finally, the (x, y, 0) is computed which denotes one corner of the installation point in the pixel space. Using four corners of the region of interest from the BIM model in equation 2 provides the RoI of the installation point in the pixel space.

2.3 Identifying the Installation of Component

Here, the objective is to use computer vision techniques to identify if the component has been installed in the pre-planned location on the modular unit, or not. Figure 4 shows the pipeline that classifies each installation region of interest as installed or not-installed.

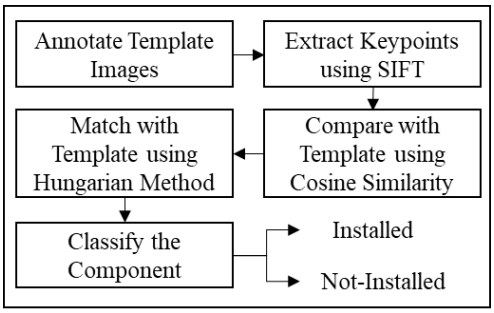

Figure 4. Identifying Installation of Component

As shown in Figure 4, a template-based matching approach is employed to compare the current state of the region of interest with template images of two possible states. Doing so, for each installation region of interest; template images of 'installed' and 'not-installed' are first annotated. Specifically, two images for each station is annotated as template from the 'installed' and 'not-installed' states. In the proposed method images are extracted and annotated from the same factory, and the same camera view. Scale Invariant Feature Transform method [23] is used to extract and describe the keypoints in the video and the templates; Cosine similarity function is used to estimate the cost of matching between the regions of the interest in the video and the template; Hungarian method is used to match the region of interest in the video with the most similar template. Finally, the component is classified as installed or not-installed based on the result of matching. Specifically, if the cost of matching between the image and the 'installed' template is less than the cost of matching between the image and the 'not-installed' state then the component is classified as installed.

The next section evaluates the proposed method on the videos collected from a modular construction factory.

3 Case Study

The proposed method was evaluated using surveillance videos captured from a modular construction factory in the US. Collected videos include five days of 12-hour shifts in a single workstation. In these videos, modular units move from one station to the next as components such as walls, and bath pods are installed. The goal of this implementation is to evaluate the proposed method by identifying the installation of a bath pod unit on a modular unit in a single workstation, and comparing the results to manual monitoring of videos.

3.1 Module Detection

Here, the modular units were detected and segmented by training the segmentation algorithm on the created dataset. Using the collected videos, an image dataset comprising 200 train and 60 test instances of modular units was annotated. Both sets were from the same factory, however, the test set was created using a camera view similar to the one in Figure 5, while the train set was created using other cameras covering other stations further down the production line. The appearance of these modules differed from the ones in the test set as the components such as walls were being installed. Next, the Mask R-CNN object segmentation algorithm with Resnet-50 backbone was pre-trained on COCO dataset using PyTorch framework. The last layer of the network was fine-tuned on the training dataset for 40 epochs, with an initial learning rate of 0.001, a momentum of 0.9, and a decay of 0.0001, using an RTX1080 GPU for 0.5 hour. Validation of the model on the test set resulted in average precision of 0.75 when Intersection over Union (IoU) parameter was set to 0.75. Figure 5 shows an example image where four modular units are detected with a bounding box and segmented with different colors. In this figure, workstation three is annotated with color purple. Figure 6 shows the annotated ground truth of this image.

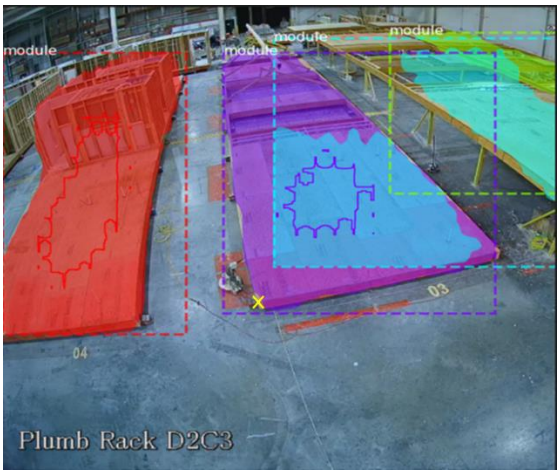

Figure 5. Example instance segmentation result

Figure 6. Ground truth annotated with yellow line

Finally, one point from each detected instance is picked as a reference at the lowest point of the detection instance. This point is annotated with a yellow cross in Figure 5. The radial distortion of the camera was disregarded, since the both source and target datasets are capture from similar cameras, however, rectification of the lens distortion can improve the performance of the detection algorithm which was left for future work.

3.2 Installation Region Identification

Here, the projection matrix was estimated and the installation region of interest was projected on the video. As shown in Figure 7, four corresponding corner points on the video are picked manually.

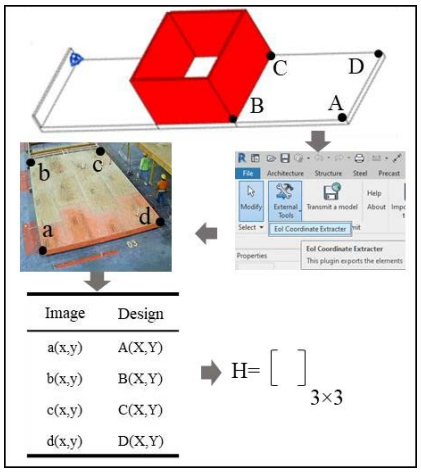

Figure 7. Extracting the corresponding points using the designed plugin

As shown in Figure 7, the corresponding points are extracted automatically using the designed BIM plugin in Revit. Here, the meta-data of the 3D module is augmented with a 'type' custom field. The designed plugin performs a search through all elements in the model filter by 'FilteredElementCollector' constructor to identify the element with type 'floor', and extracts the corner coordinates of the element and stores the data in an excel sheet. These coordinates are then used to estimate the projection matrix.

Finally, the grid and the location of the installation region is projected from the BIM model onto the video. Figure 8 shows the projected axis grids with smaller red dots and the projected boundary of the location of the installation component with purple color.

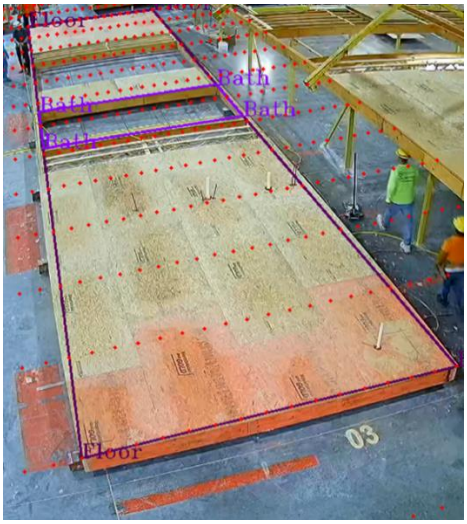

Figure 8. Projected installation location on the detected instance of modular unit

3.3 Identifying the Installation of Component

Here, the installation of a bath pod unit is monitored and identified using the proposed computer vision-based method. Figure 9 shows an example result for matching the identified installation region of interest in the detected modular unit in workstation three, with the template.

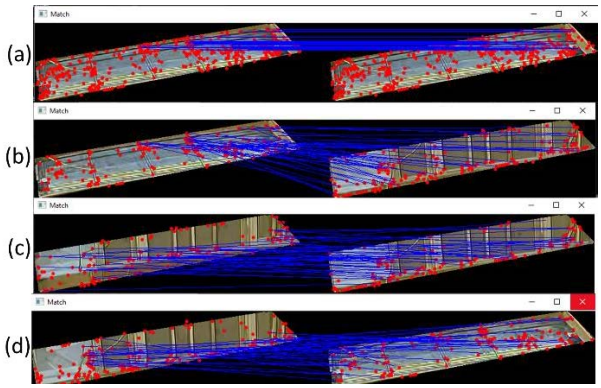

Figure 9. Matching template RoIs of the installation region on the left column with query

RoIs of the installation region on the right column: (a) matching an empty template with an empty query, (b) matching an empty template with an occupied query, (c) matching an occupied template with an occupied query, (d) matching an occupied template with an empty query

As shown in Figure 9, the 'installed' and 'not-installed' video images on the left column are matched with the templates on the right via parallel blue lines indicating strong matches. Figure 10 shows a qualitative assessment of the proposed method for identifying the installation for the bath pod unit on the video.

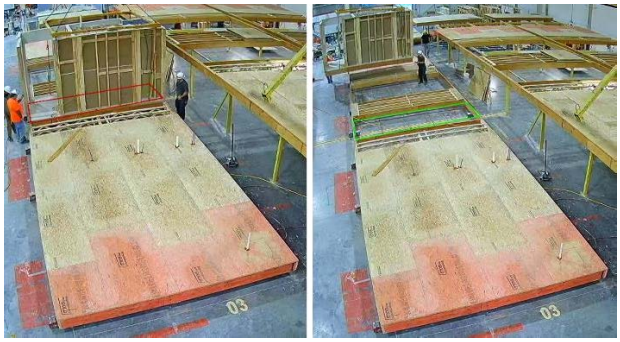

Figure 10. Installation of bath pod, as detected in the pre-planned location, on the modular unit

In Figure 10, on the right, the bath pod unit has not yet been installed in the pre-planned region which is annotated with color green. Figure 11 shows a quantitative assessment of the classification for the region of interest. Here, the same test dataset used for the module detection step, comprising 60 instances for the module, was used to evaluate the performance of the proposed installation identification method.

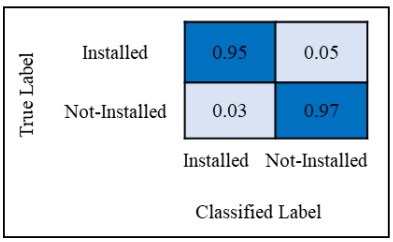

Figure 11. The evaluation performance of SIFT-based classification for region of interest

As shown in the confusion matrix in figure 11, the proposed method was able to correctly identify the installation of the bath pod on the planned location in the detected modular unit with 96% accuracy.

Analysis of the failed cases reveals that the proposed method is sensitive to occlusions caused by the presence of workers in the evaluation dataset, which lead to false positives and false negatives. Additionally, movement of the gantry crane and transportation of components such as walls, created partial occlusions and resulted in false positives due to the similarities between the bath pod and the side walls. To mitigate these temporary occlusions, median-smoothing functions can be used to classify the installation based on the median of the neighboring frames.

4 Conclusions and Future Work

This research proposed a method to monitor the progress of assembly inside modular construction factories. The proposed method relies on integration of hand-crafted and deep learning-based computer vision with building information modeling. Specifically, object detection and instance segmentation were used to precisely locate the modular units as they move from one station to the next; building information modeling and principles from projective geometry were used to identify the location of installation, based on the design, and SIFT template-based matching was used to identify the installation of the component in the pre-planned location. The proposed method was successfully evaluated on surveillance videos of a modular construction factory in the U.S. The major limitation of the proposed method is related to cases with high visual occlusion, such as scenes where the wall element was installed earlier and occluded the bath pod region of interest in the back, and manually calculating the homography matrix. To improve the proposed method, and overcome this limitation the future work will focus on improving the performance of segmentation using data augmentation and deep learning-based boundary refinement models; synthesizing the camera view using the BIM environment; projecting the 3D region of interest on the video scene, and classifying the region of interest using deep learning-based Siamese network. Furthermore, future work will consider automatically computing the homography matrix by extracting the points from the masked image without the need for user involvement.

5 Acknowledgement

This article was developed based upon funding from the Alliance for Sustainable Energy, LLC, Managing and Operating Contractor for the National Renewable Energy Laboratory for the U.S. Department of Energy. The authors also gratefully acknowledge Volumetric Building Companies for their help with data collection for this research.

5.1 References

[1] M. Arashpour and R. Wake, "Autonomous production tracking for augmenting output in off-site construction," *Autom. Constr.*, p. 9, 2015.

[2] I. Y. Wuni and G. Q. P. Shen, "Holistic Review and Conceptual Framework for the Drivers of Offsite Construction: A Total Interpretive Structural Modeling Approach," *Buildings*, vol. 9, no. 5, p. 117, May 2019, doi: 10.3390/buildings9050117.

[3] B. M. Tehrani, C. G. Ozmerdiven, and A. Alwisy, "A Decision Support System for the Integration of Robotics in Offsite Construction," in *Construction Research Congress 2022*, Arlington, Virginia, Mar. 2022, pp. 849–858. doi: 10.1061/9780784483961.089.

[4] B. M. Tehrani, S. BuHamdan, and A. Alwisy, "Robotics in industrialized construction: an activity-based ranking system for assembly manufacturing tasks," *Eng. Constr. Archit. Manag.*, Dec. 2022, doi: 10.1108/ECAM-02-2022-0143.

[5] M. S. Altaf, A. Bouferguene, H. Liu, M. Al-Hussein, and H. Yu, "Integrated production planning and control system for a panelized home prefabrication facility using simulation and RFID," *Autom. Constr.*, vol. 85, pp. 369–383, Jan. 2018, doi: 10.1016/j.autcon.2017.09.009.

[6] A. Varyani, A. Jalilvand-Nejad, and P. Fattahi, "Determining the optimum production quantity in three-echelon production system with stochastic demand," *Int. J. Adv. Manuf. Technol.*, vol. 72, no. 1–4, pp. 119–133, Apr. 2014, doi: 10.1007/s00170-014-5621-1.

[7] I. Y. Wuni and G. Q. Shen, "Risks identification and allocation in the supply chain of modular integrated construction (MiC)," *Modul. Offsite Constr. MOC Summit Proc.*, pp. 189–197, 2019.

[8] Q. Chang, J. Ni, P. Bandyopadhyay, S. Biller, and G. Xiao, "Supervisory Factory Control Based on Real-Time Production Feedback," *J. Manuf. Sci. Eng.*, vol. 129, no. 3, pp. 653–660, Jun. 2007, doi: 10.1115/1.2673666.

[9] J. Gong and C. H. Caldas, "Computer vision-based video interpretation model for automated productivity analysis of construction operations," *J. Comput. Civ. Eng.*, vol. 24, no. 3, pp. 252–263, 2009.

[10] A. Pal Singh Bhatia, S. Han, O. Moselhi, Z. Lei, and C. Raimondi, "Data Analytics of Production Cycle Time for Offsite Construction Projects," *Modul. Offsite Constr. MOC Summit Proc.*, pp. 25–32, May 2019, doi: 10.29173/mocs73.

[11] K. M. Rashid and J. Louis, "Activity identification in modular construction using audio signals and machine learning," *Autom. Constr.*, vol. 119, p. 103361, Nov. 2020, doi: 10.1016/j.autcon.2020.103361.

[12] K. M. Rashid and J. Louis, "Automated Active and Idle Time Measurement in Modular Construction Factory Using Inertial Measurement Unit and Deep Learning for Dynamic Simulation Input," in *2021 Winter Simulation Conference (WSC)*, Phoenix, AZ, USA, Dec. 2021, pp. 1–8. doi: 10.1109/WSC52266.2021.9715446.

[13] M. Ahmed, C. T. Haas, and R. Haas, "Using digital photogrammetry for pipe-works progress tracking [1] This paper is one of a selection of papers in this Special Issue on Construction Engineering and Management.," *Can. J. Civ. Eng.*, vol. 39, no. 9, pp. 1062–1071, Sep. 2012, doi: 10.1139/l2012-055.

[14] B. F. Spencer, V. Hoskere, and Y. Narazaki, "Advances in Computer Vision-Based Civil Infrastructure Inspection and Monitoring," *Engineering*, vol. 5, no. 2, pp. 199–222, Apr. 2019, doi: 10.1016/j.eng.2018.11.030.

[15] R. Panahi, J. Louis, N. Aziere, A. Podder, and C. Swanson, *Identifying Modular Construction Worker Tasks Using Computer Vision*. 2021.

[16] B. Xiao, H. Xiao, J. Wang, and Y. Chen, "Vision-based method for tracking workers by integrating deep learning instance segmentation in off-site construction," *Autom. Constr.*, vol. 136, p. 104148, Apr. 2022, doi: 10.1016/j.autcon.2022.104148.

[17] P. Martinez, B. Barkokebas, F. Hamzeh, M. Al-Hussein, and R. Ahmad, "A vision-based approach for automatic progress tracking of floor paneling in offsite construction facilities," *Autom. Constr.*, vol. 125, p. 103620, May 2021, doi: 10.1016/j.autcon.2021.103620.

[18] W. Chu, S. Han, X. Luo, and Z. Zhu, "Monocular Vision–Based Framework for Biomechanical Analysis or Ergonomic Posture Assessment in Modular Construction," *J. Comput. Civ. Eng.*, vol. 34, no. 4, p. 04020018, Jul. 2020, doi: 10.1061/(ASCE)CP.1943-5487.0000897.

[19] Z. Zheng, Z. Zhang, and W. Pan, "Virtual prototyping- and transfer learning-enabled module detection for modular integrated construction," *Autom. Constr.*, vol. 120, p. 103387, Dec. 2020, doi: 10.1016/j.autcon.2020.103387.

[20] K. Park, S. Ergan, and C. Feng, "Towards Intelligent Agents to Assist in Modular Construction: Evaluation of Datasets Generated in Virtual Environments for AI training," presented at the 38th International Symposium on Automation and Robotics in Construction, Dubai, UAE, Nov. 2021. doi: 10.22260/ISARC2021/0046.

[21] K. Park and S. Ergan, "Toward Intelligent Agents to Detect Work Pieces and Processes in Modular Construction: An Approach to Generate Synthetic Training Data," in *Construction Research Congress*

2022, Arlington, Virginia, Mar. 2022, pp. 802–811. doi: 10.1061/9780784483961.084.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99. Accessed: Feb. 13, 2020. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf

[23] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.