A vision-based approach for automatic progress tracking of floor paneling in offsite construction facilities

Pablo Martinez a,b, Beda Barkokebas a, Farook Hamzeh a, Mohamed Al-Hussein a, Rafiq Ahmad b

a Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada
b Laboratory of Intelligent Manufacturing, Design and Automation (LIMDA), Department of Mechanical Engineering, University of Alberta, Edmonton, Canada
*Corresponding author. E-mail address: rafiq.ahmad@ualberta.ca

Abstract
Offsite construction is an approach focused on moving construction tasks from traditional jobsites to manufacturing facilities. Improved productivity of construction tasks is paramount in terms of competitiveness and is achieved through the continuous improvement of operations and planning, which often relies on historical data obtained from previous projects. Despite being a common practice, current methods, such as time studies, are not able to capture the changing scenarios resulting from improvements to production. This paper presents a novel approach to automatically detect and track the progress of construction operations by applying a method that combines deep learning algorithms and finite state machines to existing footage captured by closed-circuit television (CCTV) security cameras. Applied in the context of floor panel manufacturing stations, the proposed method examines entire production days recorded by CCTV cameras, while providing the durations of each task, its required resources, and the task efficiency per panel with high accuracy.

Keywords: offsite construction; construction automation; computer vision; productivity; machine learning; task efficiency.

1 Introduction

1.1 Background

Offsite production is an increasingly popular approach in construction that relocates most on-site operations to a more controlled factory environment. Offsite construction facilitates a hybrid approach that can be described as a series of construction operations in an assembly line, which offers several advantages over traditional construction such as mass customization, increased productivity and quality, improved safety and health for construction workers, minimization of construction waste, and shorter delivery times [1]. The competitive advantage comes from the adoption of automation, innovative facility layouts, and broad adoption of information technology. The performance of offsite facilities depends on labor and production planning because poor planning can translate into bottlenecks on the production line and an increase in production costs [2]. Consequently, delays create a gap between planned production and actual output, which prevents offsite practitioners from meeting their scheduled commitments. These delays are associated with inaccurate productivity metrics for floor panel production due to the high degree of variability in the cycle time at the workstations resulting from variations in the panel design specifications and due to the lack of accurate data collected from the production line [3]. Despite significant performance improvements at offsite construction facilities, studies aimed at productivity improvement are based on employing manual observations to monitor construction activities, a method that is error-prone and often provides wide ranging of results [4]. Furthermore, manual work monitoring limits productivity enhancements in offsite construction, since it only captures the state of production at the time it is performed and is not sustainable for long-term planning of operations due to the ever-changing nature of improvements performed at the facility.

Systematic monitoring of construction operations can bring an immediate awareness of task-specific issues [5,6]. It provides stakeholders with necessary information related to scheduling, costs, productivity, and resource utilization, that quickly supports project control decision-making [7]. Presently, a large number of data collecting technologies are used for progress tracking, from radio-frequency identification (RFID) to aerial photogrammetry [8]. The selection of the correct technology, however, depends heavily on the project requirements and the level of readiness of the required data [9]. As a type of easily captured and widely spread media, images and videos have become popular in the architecture, engineering and construction
Applying vision-based technologies to analyze the recorded images and videos automatically has drawn much attention from practitioners and from academia [10]. Several interdisciplinary works have enabled the measuring, detecting and tracking of objects, i.e. equipment and/or workers, which play a critical role in construction performance monitoring applications [11]. However, most studies focus on on-site activities [12], which means the proposed solutions are hardly applicable to the offsite construction sector [9]. Indeed, Arashpour et al. [1] indicate the deficiency of on-site production tracking systems for use in offsite construction facilities due to the fundamental ways the offsite approach differs from the on-site approach, such as the accelerated pace at which the trades work in manufacturing workstations and the long-term production targets (e.g., daily, weekly, or even monthly quotas), thus, requiring short- and long-term production tracking at the same time. This information is required to monitor production deviations throughout the day and provide real-time information so management can address bottlenecks in a timely manner while being aware that these deviations will have an impact to some degree on production in the following days. This requires tracking systems that provide consistent, reliable, and accurate data for short and long periods of time.

Whereas site-built methodologies focus on identifying the key resources among the crowded job site to determine the current activity or scheduled task, i.e. workers/crews [13–15] or heavy equipment [16], offsite facilities consist of workstations, each with limited and sequenced activities that should be easier to identify. However, compared to construction sites where cameras are temporarily located and oriented specifically for the monitoring task at hand, cameras are already installed in offsite facilities for closed-circuit television (CCTV) security footage of the facility. Such cameras typically provide low-resolution videos to facilitate long-term storage. In fact, low-resolution cameras are recommended for workplace surveillance due to privacy issues and workers’ mental health [17,18]. However, the reported performance levels of state-of-the-art activity monitoring methods using such cameras are below 50% in terms of accuracy [19]. To overcome the low resolution environment and enable automatic monitoring of tasks in offsite construction facilities, this paper proposes a hybrid method that combines deep learning algorithms and virtual finite state machines (VFSM). This approach uses the accurate vision-based detection and classification of novel deep learning algorithms, such as the Faster Region-based Convolutional Neural Network (R-CNN), along with robust computationally modeled transitions between sequences of events, for which VFSM is a widely used technique in the robotics field [20]. Moreover, this approach provides a novel method to capture productivity-related metrics, such as duration and number of workers per task, in order to address and enhance the efficiency of offsite construction operations. The methodology is tested and finally validated using video footage from a floor panel manufacturing workstation in an offsite construction facility.

1.2 Literature review

1.2.1 Process improvement in offsite construction

Offsite construction is often associated with shortened schedules and increased productivity in different parts of the world [21,22], allowing on-site construction work, i.e. foundations, to be performed concurrently with the fabrication of the project’s structure in the form of panels or volumetric modules. The increased productivity results from both external and internal factors: (1) a controlled environment immune to the influence of weather conditions [23], and (2) higher productivity in construction operations due to process improvement studies conducted at the factory. Indeed, offsite construction provides a more suitable environment for data collection, which allows practitioners to collect more reliable data due to the controlled environment and reduced variability of motions from workers [24]. Through interviews and observations, it is possible to reduce production hours considerably by creating a culture of continuous improvement and identification of value and non-value added tasks in the overall process [25], while applying other methods to predict other aspects of production. Sanderg and Bildsten [26] identify overproduction, waiting, and needless movements as the most critical waste activities in offsite construction production because they render workstations idle and non-productive.
Different techniques are applied to address value in offsite construction tasks such as value stream mapping, which identifies opportunities for improvement to reduce production lead times [27], and fuzzy analytic hierarchy processes, which rank expert opinions to estimate the risks associated with delays and cost overruns from the project’s initial baseline [28]. From these analyses, several methods are employed to improve construction tasks and improve productivity at the manufacturing facility. Mullens and Kelley III reduces the task’s cycle time by eliminating non-value-added motions at workstations and by constantly monitoring production output [29]. A dependency structure matrix is applied to prevent bottlenecks, rework, and congestion at stations to reduce the number of main activities by 30% through the observation of construction tasks and their inter-dependence [30]. The impact of multi-skilled labor in offsite construction facilities is also addressed, taking into consideration various strategies for its implementation and quantitative indicators pertaining to productivity and cost [31]. Other studies consider a holistic approach to quantify the economic, social, and environmental aspects of production to suggest improvements and maintain a continuous improvement culture in the company [25].

Despite the relevant improvements recommended in the aforementioned publications, all input data is collected in the form of time studies of construction tasks, usually taken manually, which is an error-prone, monotonous, tedious, and time-consuming process. Moreover, manual time studies fail to address the impact of changes on the production line, thus becoming obsolete after a meaningful improvement of a task [32], requiring a new effort to collect updated data for further improvement. This poses as a significant barrier to continuous improvement in offsite construction facilities over the long term without significant investment and commitment from the companies. In order to address this issue, the use of real-time tracking systems, such as RFID and barcodes, are proposed to provide real-time feedback to upper management regarding inventory and production while maintaining a permanent method of communication between production and process improvement teams [33].

Presently, applications using real-time tracking systems are scarcely employed in offsite construction, especially in the context of managing production lines. Nevertheless, cost savings and increased productivity are already observed in offsite construction facilities when using barcodes to track material flow and inventory [34]. Moreover, production schedules are optimized using RFID tags when taking into consideration wall panels’ attributes, job sequencing, and constructability aspects from production [35]. RFID systems are also applied in quality and production tracking, providing a system for proactive process improvement on production lines [36]. Despite the benefits involved in this approach, doubts regarding the implementation of real-time tracking systems are raised due to the inherent trade-off between the cost to implement these technologies and any quantifiable gains in productivity. Anderl and Fleischer [37] argue the initial investment in hardware limits a broader application of these systems, thus necessitating alternative approaches to implementation. In order to provide an alternative solution for the implementation of real-time tracking systems in offsite construction facilities, this paper proposes the use of computer vision techniques to track process and idle times from existing CCTV security footage captured by cameras installed to monitor a production line.

1.2.2 Computer vision applied to work monitoring

Despite growing interest and relevant work in the past twenty years, there is a bias in research literature towards traditional construction over offsite construction involving the application of computer vision [38]. Studies involving the application of computer vision in offsite construction are scarce across several research areas. Martinez et al. propose a system to perform real-time quality inspections of light steel-framed panels by comparing the image of the assembled panel with its intended design [39]. Moreover, algorithms are developed to detect the work progress of wall panels during construction, taking into consideration different components of walls at different stages [40]. Planning and final installation of manufactured elements at their designated sites are also the subject of several studies, such as the combination of various information systems and computer vision for the installation of structural elements [41], and the identification of images to detect module lifting tasks for the final assembly of high-rise buildings [42].
The application of computer vision in traditional construction environments has advanced work monitoring through the automatic detection of labor and work progress in a rapid and accurate manner [43]. Construction tasks and finished products are similar in the context of both traditional and offsite construction, which provides interesting insights for work monitoring in the latter scenario. However, the significant differences in the physical spaces must be taken into consideration for accurate analysis and are addressed in this paper. A significant challenge for image processing in traditional construction scenarios is the constant change of reference from viewpoints and the Region of Interest (RoI), the area of the image from which information is extracted, due to objects obstructing the camera on-site [44]. Moreover, constant changes in illumination pose a significant challenge in terms of the accuracy of computer vision algorithms [45], while the installation of permanent tracking devices on traditional construction sites poses challenges, such as the need for electricity and concerns about theft [46]. Overall, from a computer vision perspective, these challenges are significantly reduced in an offsite construction facility: (1) workstations are fixed and worker motions are more predictable according to the factory layout; 2) lighting conditions are stabler throughout the day, and sun glare and shadowing effects are minimum; and, 3) cameras are permanently installed at fixed spots, free of obstructions due to their original goal (security), meaning they would not require relocation nor would there be any concerns about theft.

Several studies support apply vector machines (SVM) and other computer vision algorithms, such as EDLines and line segment detectors (LSD), to identify building elements (walls, equipment, etc.) and their progression during the project based on the images collected on-site in order to identify the progress of activities during construction [47–49]. Despite being a good solution for identifying the progress of a project, non-permanent data collection of images inhibits further productivity analysis at the activity level since the number of workers and time spent on each activity is not recorded. Hence, to monitor the work and productivity of construction tasks at an offsite construction facility, computer vision algorithms must be able to track workers and recognize the task being performed through video to monitor both progress and productivity despite the significant extra effort to process videos instead of static images. Gong and Caldas [49] describe the main three approaches for automated productivity measurement using video images: (1) trajectory recognition and tracking of resources, (2) movement detection of construction resources, and (3) recognition of worker’s gestures. Trajectory recognition may not be very applicable in the context of automated productivity measurement in offsite construction since the tasks are contained within each workstation, and internal trajectories of workers are not without purpose. Therefore, movement detection of equipment (e.g., cranes, multi-function bridges, etc.) and recognition of worker’s gestures are more suitable approaches for the implementation of computer vision algorithms to measure productivity in offsite construction automatically.

Despite the positive results, Luo et al. [54] argue that the changes in background, which can include image obstruction, illumination variations, and viewpoint changes, combined with multiple interactions and group activities reduce the accuracy of the proposed methods, thus suggesting the application of convolutional neural networks (CNN) to overcome these issues and automatically recognize workers’ activities. CNNs can accommodate complex tasks while simultaneously capturing static, short-term, and long-term motions in a video from large datasets in a timely manner [52]. In fact, CNN architectures have been identified as the main support for robust monitoring of construction resources [53,54], and have been applied in a number of different areas, such as safety [55], progress monitoring [56], and resource localization [57]. This is because neural networks, including CNNs, have superior capabilities in terms of representing complex relationships between inputs, i.e., images, and desired outputs, i.e., resources (objects). Most recent frameworks with the aim of monitoring resources on job sites include CNNs for the initial object detection [58,59].

Continuous monitoring of construction projects allows project managers to evaluate the operational efficiency of their processes and input resources (i.e., crew production rates), determine risk factors that can cause delays or safety accidents, and analyze current construction progress [60]. R-CNNs have been recently used to detect various types of construction objects, including workers and equipment to enable such close monitoring of operations. Fang et al. used such approach to provide efficient monitoring of heavy equipment and operators [19]. Kim et al. trained a similar neural network to monitor the construction productivity in an earthmoving process [61]. Luo et al. built an activity recognition method, based on CNNs,
to detect multiple construction resources and interpret their spatial relationships in order to estimate the operational efficiency of the construction processes and resources used [62]. Similarly, job site crowdness and the possibility of potential dangerous physical interferences between resources have been proposed to be monitored by using R-CNNs [57,63]. Overall, deep learning algorithms have shown great performance on vision-based construction monitoring. For this study, the authors follow the current trend of incorporating CNN architectures for the purpose of identifying resources within offsite construction facilities and providing robust data to monitor repetitive tasks.

### 1.2.3 Finite state machines applied to construction

A finite state machine (FSM) is commonly used as a control strategy for physical devices and, as a mathematical construct, is used to approximate a broad range of physical or abstract phenomena [64]. A FSM aims to control a sequence of actions depending on certain triggering events or rule sets, where the number of actions or states is a predefined finite list. The FSM changes its state when a condition is satisfied through a state transition. A FSM can be mathematically defined as a quintuple \((K,S,S_0,C,S_f)\): \((K)\) is the set of events, which acts as an input of the FSM; \((S)\) is the predefined, finite, and non-empty set of states; \((C)\) is the state-transition functions, \(C = K \times S\); and \((S_0)\) and \((S_f)\) are the initial and final states, respectively, which are subsets of \((S)\). A virtual finite state machine (VFSM) is a finite state machine that is completely defined and executable in a virtual environment. A VFSM offers an alternative to rule-based algorithms to describe the behavior of a system or process without any physical interference or action [65]. VFSMs have been considered a more robust and transparent approach to represent dynamic systems in comparison to rule-based systems, especially considering the strength of having stable states that minimize the potential of erroneous identification. Targeting the automation of any process, the FSM has played a key role in the modeling and testing of autonomous equipment for the past several decades [66,67].

In construction, however, the integration of state machines in autonomous or automatic construction processes is rare. A few recent examples can be found in the autonomous control of critical equipment, i.e., slurry shield tunnel boring machinery [68], or modeling of construction equipment activities, i.e., trucks and excavators during earthmoving operations [69]. In general, modeling through state machines has been overlooked by the academic community in regard to construction activities. Researchers in the construction field prefer to model construction operations using discrete-event simulation (DES) models, which are especially popular to monitor construction operations on job sites [70]. While both methods rely on consistent and accurate input data to build a correct model, DES and VFSM yield different benefits in the long term. DES provides after-de-facto modeling and simulation of different scenarios to assist in determining which future actions will improve productivity. Meanwhile, VFSMs can be executed and provide monitoring data in real-time (e.g. task durations) while storing this information to address the task efficiency after it is finished or compare its data with similar cases. From a construction management perspective, DES modeling synchronizes adequately with the project-centered, schedule and cost-driven construction mentality [71], and justifies the academic focus on this modeling approach. VFSM, on the other hand, is able to instantly provide data that can be analyzed to determine the productivity and reliability of operations (e.g. task efficiency) in real-time based on the system’s change of states which are not, in most cases, of high interest to construction managers [72]. Moreover, offsite facilities provide an interesting hybrid environment between manufacturing and construction that can benefit from both approaches. A few frameworks that focus on employing DES in offsite construction already exist in the literature: for example, to plan offsite construction projects [73], or to model inventory in panelized construction facilities [74]. This paper demonstrates the potential of VFSM as a modeling tool for construction activities, namely in the offsite construction industry.

### 2 Methodology

The proposed approach adopts a design science research methodology to track the progress of offsite construction operations using computer vision techniques. Koskela [75] differentiates design research science from other types of methodologies as that which develops an artifact: something that is useful and improves the problem at the identified and explained research
The process of developing an artifact consists of a rigorous procedure of identifying gaps in the literature, and developing the artifact and its evaluation methods in a structured and replicable manner while clearly communicating its outputs [76]. The artifact developed in this paper consists of a vision-based approach to track the work progress in offsite construction operations in a timely manner and with a high level of accuracy from low-resolution images collected from CCTV footage. In possession of this artifact, it is possible to track work progress from a large sample of panels without being restricted to few examples collected during time studies, while making use of existing hardware (CCTV cameras) and minimizing the initial investment required for a traditional approach such as the purchase of high-resolution cameras. This methodology has been applied equally as successfully in the development of advanced productivity metrics and resource planning for construction as it was in the development of computer vision frameworks focused on improving safety operations in construction [77–80]. The methods applied in this research are demonstrated in Error! Reference source not found. and are divided into three stages: (1) the descriptive stage, which consists of the identification of task motions and their sequence contained in camera footage from field and case studies with the aim of describing the recorded tasks; (2) the modeling stage, which includes this information in the proposed approach as states in the VFSM engine for its identification using R-CNN techniques; and (3) the evaluation/testing stage, which consists of the evaluation and testing of the proposed approach. Taking into consideration the value-added and the non-value-added tasks in the process and the expert knowledge required to describe the sequence and operations involved, the proposed approach provides information regarding the panels and the overall production duration and performance.

Figure 1. Overview of proposed method.

The approach presented in this paper aims to automatically track the progress of construction tasks in offsite facilities using low-resolution videos from CCTV footage. The proposed approach uses novel image processing techniques to filter, detect, and classify resources from the video, and computational techniques to strengthen the tasks’ transitions that determine the monitoring results. An overview of the approach is illustrated in Figure 2, and the details regarding its implementation in the presented case study are provided in the following sections.
First, the task to be monitored needs to be pre-analyzed to determine the resources to be detected using the deep learning algorithms. Then, the VFSM states are defined to represent the events of the task, i.e., placing and securing the joists in a floor panel manufacturing station, while the transitions are linked to the detection or non-detection of the resources in the area. As such, the current state determined by the VFSM corresponds to the current progress of the task. Finally, the sequence of states, as well as the detection results and video timestamps, can be exported for offline analysis.

3 Case study: floor assembly manufacturing

3.1 Virtual finite state machine for floor assembly manufacturing

For this study, VFSM is used to model the required tasks that occur at a floor panel assembly workstation in an offsite manufacturing facility. Currently, at the station under study, floor panel assemblies are manufactured using a semi-automated six-step process. First, operators manually place the floor joists over a flat surface following pre-designed shop drawings. Then, an automated process sets a thin layer of glue on top of each joist. Once the gluing process is finished, operators place the sheathing boards on top of the joists. Once the glue has dried, the excess sheathing is automatically cut and posteriorly manually removed. Finally, the finalized floor panel is lifted with a crane out of the station for another panel to be manufactured. The states, $S$, of the VFSM are directly associated with the subtasks that occur during the process of manufacturing a floor panel. The state set is defined as $S = \{ \text{Idle, Joists Placement, Auto Gluing, Sheathing Placing, Auto Stapling & Cutting, Cleaning Inspection & Lift Preparation, Crane Lifting} \}$. Then, the state transitions, $C$, represent the visible variations in activity at the station, targeting differences in the resources needed for manual or automated subtasks. These transitions are directly controlled by a select number of variables that are continuously updated based on the output of the R-CNN. Table 1 presents the variables for the selected case study. Note that knowing that the state transitions come from R-CNN detections, it is assumed that workers that are detected within the boundaries of the station are there to work. This assumption is based on the predictability of worker walking patterns on offsite construction facilities [81].

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
<th>Type</th>
<th>Initial Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workers</td>
<td>Represents the number of operators detected working in the selected working area.</td>
<td>Integer</td>
<td>0</td>
</tr>
<tr>
<td>Crane</td>
<td>Establishes whether the lifting crane has been detected in the selected working area.</td>
<td>Binary</td>
<td>0</td>
</tr>
</tbody>
</table>
Machine Establishes whether the multi-function bridge machine has been detected in the selected working area. Binary 0
prevState Stores the identification number of the previous state. Integer 0

Figure 3 shows the states and transitions in the VFSM. In total, there are 7 states and 15 state transitions. The state ‘Idle’ represents a state in which the station is completely inactive and acts as the starting state, (\(S_0\)). No final state is defined in this particular case, as the process is cyclical: as soon as a floor panel is finished, a new one is starting. As deterministic state machines, the VSFM leads to a single state when given a particular input from the R-CNN.

![Virtual finite state machine for floor panel manufacturing operations in offsite construction facilities.](image)

**3.2 Region-based convolutional neural network (R-CNN)**

This study researches the use of region-based convolutional neural network (R-CNN) to provide automatic detection and classification of resources used in a specific offsite construction manufacturing task. By taking images and object proposals from selective searches as inputs, R-CNNs use convolutional neural networks (CNNs) to extract features, and then locate and classify objects based on the initial search parameters. The following subsections detail the dataset used to train the R-CNN, its architecture, and the training and validation results for the proposed case study.

**3.2.1 Dataset**

To develop a database containing information regarding the tasks at the floor panel manufacturing station, 1069 images (with a resolution of 640x480 pixels) are collected from the CCTV system of an offsite facility in Edmonton, Canada. The selected images are taken randomly from several days of video footage during working hours from the same CCTV camera used in...
this study. To annotate the images with labels for both the inspection results and the coordinates of their corresponding bounding boxes, the software environment, MATLAB, is used to specify them manually. For the case study presented, three labels are used to determine the resources that need to be detected: workers (labeled ‘worker’), the multi-function bridge machine (labeled ‘machine’), and the crane lifting system (labeled ‘crane’). Labeling is determined, it should be noted, based on the resources that trigger events in the finite state machine that is defined by the task at hand. An example of the images used in the dataset can be found in Figure 4.

![Sample images of the CCTV footage used in the training dataset with the corresponding labels and bounding boxes (cyan: worker, magenta: machine, and orange: crane).](image)

**3.2.2 R-CNN architecture**

The Faster R-CNN designed for use in this study is based on a pre-trained ResNet-50 architecture. The choice of the CNN architecture is based on ease of access and availability, thus, other architectures such as YOLO or MobileNet could have been used to obtain similar results. Transfer learning, it should be noted, has proven to be effective in reducing training time and computational demand, and in providing accurate results, even with small datasets [82]. By skipping connections within the network, ResNet enhances the detection of smaller objects in images, such as the workers for this study. To adapt it to the context of this study, the last three layers of the ResNet-50 are substituted by a fully connected layer, a SoftMax layer, and a classification layer, themselves connected by average pooling. Furthermore, additional convolutional layers and activation layers are connected to the feature extraction layer of the ResNet-50, themselves connected to the region proposal layers. These final layers drastically improve the bounding box regression, improving accuracy and supporting a clearer classification of the detected object. All the added components are then retrained for the purpose of detection and classification of the aforementioned labels.

**3.2.3 Training and validation results**

To generate a training and validation dataset, images are randomly selected from the labeled images such that each of the label types—workers, machine, and crane—represent at least 30% of the images contained in the validation set. The remaining images not selected are used for training the Faster R-CNN. Both training and validation are performed using the open-source Faster R-CNN library available within MATLAB2019b (computing software). The neural network is trained using stochastic gradient descent, with a momentum of 0.9, an initial learning rate of $10^{-3}$, and a batch size of 32. The training of Faster R-CNN networks is divided into four training steps: T1) initial training of the region proposal network (RPN); T2) training of the Faster R-CNN network with RPN; T3) retraining of the RPN using Faster R-CNN weights; and, T4) retraining of the Faster R-CNN with the updated RPN. The accuracy and loss function for both training and validation steps are recorded and shown in Figure 5. After 35 epochs, the accuracy of the network reaches over 92% for both the training and validation data. The Faster R-CNN reaches a competitive accuracy performance and is considered robust enough to analyze its performance in the following sections. An example of image results from the validation test is shown in Figure 6.
Figure 5. Faster R-CNN training (T4) and validation results.

Figure 6. Results from the validation dataset with corresponding bounding boxes, labels, and confidence scores.
3.3 Region of interest (RoI)

As cameras are originally placed for security purposes, a filter designed to limit the extraction of information to a specific area or region of interest (RoI) is used. As the camera is static, the same RoI is applied during the entire duration of the footage. The RoI is defined by a set of vortexes, $V = \{V_1, V_2, \ldots, V_n\}$, that are pre-determined for each camera and target station. Then, the vortexes define an area of the image in which a binary polygonal mask is applied over the image. As such, all the pixels that are not within the RoI are turned to black while the rest is left as is. For the station used in this study, Figure 7 illustrates the RoI applied in this study and the resulting images obtained. This step is especially relevant, for example, if more than one station under study is visible in the video footage used. For the footage used in this study, the top-left portion of the image shows other stations in the offsite facility, namely the joist cutting station, where operators work and other machinery is used. As operations begin in that station, detections in that area would produce extraneous, unwanted results.

![Figure 7. Region of interest for the floor panel station under study. Left: definition of vortex set. Right: resulting image after applying the binary mask used as input to the Faster R-CNN detector.](image)

4 Test results and discussion

This section aims to validate the proposed methodology by monitoring and analyzing a full eight-hour shift at a floor panel assembly station. First, the results obtained for an eight-hour shift are presented and thoroughly explained. Finally, the results and limitations of the proposed approach are discussed in depth.

4.1 Results

During the selected shift, six floor panel assemblies were scheduled to be manufactured. These assemblies vary in length, shape, number of joists, and so forth. In order to compare the results obtained through the proposed methodology, a manual analysis of the same video footage is performed. This analysis targets two metrics: duration, and man-hours required for each task. For each task, the number of workers is pre-assigned as per the schedule, as listed in Table 2.

Table 2. List of tasks in the floor panel assembly station and corresponding manpower assignment.

<table>
<thead>
<tr>
<th>Task</th>
<th>Corresponding State (VFSM)</th>
<th>Manpower Assigned [# operators]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepare, Place, and Secure Joists</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Auto-Gluing</td>
<td>2</td>
<td>0 (Automated)</td>
</tr>
<tr>
<td>Task</td>
<td>Duration</td>
<td>Man-hours</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Place Sheathing Boards</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Auto-Stapling and Cutting</td>
<td>4</td>
<td>0 (Automated)</td>
</tr>
<tr>
<td>Cleaning, Inspection, and Lift Preparation</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Crane Lifting Operation</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

Based on the current state, manual analysis determines the duration, \(d\), and man-hours, \(m\), required for the corresponding task following Equation 1:

\[
\begin{align*}
\{ d &= t_{i}^{\text{end}} - t_{i}^{0} \\
\{ m &= d \ast w_{i} \}
\end{align*}
\]  

(1)

where \((t_{i}^{0})\) and \((t_{i}^{\text{end}})\) represent the timestamps at which a task \((i)\) has started and finished, respectively, and \((w_{i})\) is the manpower pre-assigned for the task \((i)\). However, for the proposed approach, both metrics are automatically determined based on video parameters and frame-by-frame results from both the R-CNN and VFSM, as explicitly defined in Equation (2):

\[
\begin{align*}
\{ d &= n_{i} \ast f \\
\{ m &= \sum_{j} w_{j} \ast f \}
\end{align*}
\]  

(2)

where \((n_{i})\) represents the number of elapsed frames for a certain task \((i)\), \((w_{j})\) represents the number of workers (operators) detected in the frame \((j)\), and \((f)\) is the video framerate or the time elapsed between two frames introduced into the system. For the video presented, the CCTV system outputs 25 frames per second. Using both equations, the analysis of each panel assembly process at the station under study is performed.

First, the results obtained by applying the R-CNN to the studied video footage are presented. To present the results obtained by the proposed approach in a concise manner, a single floor panel is initially selected. The analyzed panel corresponds to “Floor Panel ID #3”, or in other words, the third panel manufactured during that shift. Note that three labels are defined as targets for detection and classification by the algorithm: workers, machine, and crane. Figure 8 Error! Reference source not found. illustrates the detection results for each one of the three labels. ‘Detected’ shows the results from the R-CNN, ‘Real’ is obtained by manual human inspection frame by frame, and ‘As Scheduled’ follows the rules specified in Table 2.
Figure 8. R-CNN results in the manufacturing of floor panel ID #3. Top: worker label. Bottom left: machine label. Bottom right: crane label.

For the presented results, the precision, recall, and F₁ score are computed to check the performance of the proposed R-CNN. Precision is the ratio of correctly predicted positive observations for a specific label with respect to the total predicted results. Recall or sensitivity is the ratio of correctly predicted positive observations for a specific label to the total number of observations of that label. The F₁ score is a weighted average of precision and recall that takes into consideration both positive and negative observations, and results in a metric that represents the overall accuracy of a specific label. These metrics have been used consistently in academia to measure neural network performance where a small number of positive instances for a label are present, which is the case for this study. The metrics are defined in Equations 3–5:
where \( TP \) (true positive) denotes the number of detections classified correctly, \( FP \) (false positive) denotes the number of detections classified incorrectly, and \( FN \) (false negative) denotes the number of undetected labels. These values are obtained by comparing the classification results to manual inspection results (labeled ‘Manual Video Analysis’). As such, Table 3 shows the aforementioned performance metrics for the floor panel assembly under study.

Table 3. List of R-CNN results and performance metrics for the manufacturing of floor panel ID #3.

<table>
<thead>
<tr>
<th>Label</th>
<th>R-CNN Results</th>
<th>Performance Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( TP )</td>
<td>( FP )</td>
</tr>
<tr>
<td>Worker</td>
<td>96843</td>
<td>1142</td>
</tr>
<tr>
<td>Machine</td>
<td>36276</td>
<td>129</td>
</tr>
<tr>
<td>Crane</td>
<td>1972</td>
<td>12</td>
</tr>
</tbody>
</table>

As observed, the performance of the R-CNN is satisfactory for worker and machine detection and classification, with an \( F_1 \) score of 0.932 (0.947 overall in the 8-hour shift analysis) and 0.997 (0.983 overall in the 8-hour shift analysis), respectively. This is not surprising as the detection targets are relatively large in size, and for the machine label, in particular, no comparable object can be found for which it could be mistaken. For the crane label, the calculated \( F_1 \) score is lower, 0.722 (0.745 overall in the 8-hour shift analysis), compared to the other labels. Low results for crane detection are expected due to targeted features that are much smaller and complex than for the other labels, because they are aiming at detecting lifting points on the manufactured floor panel and at detecting all the used shackles. These results are used as input for the VFSM. The sequence of states obtained from the VFSM is shown in Figure 9.
As observed, the proposed VFSM tracks, with an accuracy of 98.7%, the progress of the floor panel assembly manufacturing task, even with the low-performance R-CNN results obtained for the crane label. Accuracy, in this case, is calculated as the ratio of total time the VFSM reports the correct state (task progress) to the total manufacturing time of a floor panel assembly. The strength of state transitions of VFSM supports lower performance detectors in low-resolution environments to reach accurate task progress tracking. Nonetheless, VFSM errors in tracking task progress do occur and can be divided into two groups: 1) errors due to false negatives in worker detection that cause the VFSM to turn Idle for a small number of frames; and 2) worker detection during Idle phases, for example, due to walk-throughs, or false positives that cause the VFSM to leave the Idle state for a small number of frames. Both errors are highlighted in Figure 9. Focusing on the metrics as defined in Equations 1 and 2, the overall results are listed in Table 4.

### Table 4. Comparison of results obtained from a single floor panel assembly.

<table>
<thead>
<tr>
<th>State ID</th>
<th>Task</th>
<th>Manual Video Analysis w/ Manpower as Scheduled</th>
<th>Proposed Approach</th>
<th>Relative Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Idle</td>
<td>Duration [h] 0.3778</td>
<td>0.000</td>
<td>0.3861</td>
</tr>
<tr>
<td>1</td>
<td>Prepare, Place and Secure Joists</td>
<td>Duration [h] 0.2139</td>
<td>0.428</td>
<td>0.2083</td>
</tr>
<tr>
<td>2</td>
<td>Auto-Gluing</td>
<td>Duration [h] 0.0889</td>
<td>0.000</td>
<td>0.0861</td>
</tr>
<tr>
<td>3</td>
<td>Place Sheathing Boards</td>
<td>Duration [h] 0.1486</td>
<td>0.297</td>
<td>0.1750</td>
</tr>
<tr>
<td>4</td>
<td>Auto-Stapling and Cutting</td>
<td>Duration [h] 0.3153</td>
<td>0.000</td>
<td>0.3167</td>
</tr>
<tr>
<td>5</td>
<td>Cleaning, Inspection, and Lift Preparation</td>
<td>Duration [h] 0.1319</td>
<td>0.264</td>
<td>0.1028</td>
</tr>
<tr>
<td>6</td>
<td>Crane Lifting Operation</td>
<td>Duration [h] 0.0347</td>
<td>0.035</td>
<td>0.0389</td>
</tr>
</tbody>
</table>

In summary, duration and man-hours can be determined using the proposed approach. In terms of accuracy, the duration is calculated with an error of 0.22% (approximately 10 seconds), and man-hours are directly computed from the R-CNN worker detection rate, with an average error of 2.5%. The discrepancies between scheduled and calculated man-hours will be discussed in the following subsection. Finally, Table 5 summarizes the results obtained using the proposed approach and compares it to the manual analysis.

### Table 5. Comparison of results obtained from the 8-hour shift video footage of the floor paneling assembly station.

<table>
<thead>
<tr>
<th>Floor Panel ID</th>
<th>Manual Video Analysis w/ Manpower as Scheduled</th>
<th>Proposed Approach</th>
<th>Relative Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Start Time 00:12:50</td>
<td>End Time 01:27:15</td>
<td>Duration [h] 1.2403</td>
</tr>
<tr>
<td>2</td>
<td>Start Time 01:27:15</td>
<td>End Time 02:04:25</td>
<td>Duration [h] 0.6194</td>
</tr>
<tr>
<td>3</td>
<td>Start Time 02:04:25</td>
<td>End Time 03:23:05</td>
<td>Duration [h] 1.3110</td>
</tr>
<tr>
<td>4</td>
<td>Start Time 03:23:05</td>
<td>End Time 04:43:35</td>
<td>Duration [h] 1.3420</td>
</tr>
<tr>
<td>5</td>
<td>Start Time 04:43:35</td>
<td>End Time 06:51:20</td>
<td>Duration [h] 2.1292</td>
</tr>
<tr>
<td>6</td>
<td>Start Time 06:51:20</td>
<td>End Time 07:54:40</td>
<td>Duration [h] 1.0694</td>
</tr>
</tbody>
</table>

Overall, the average error for the duration of floor panel manufacturing is 0.74% or approximately 26 seconds with a standard deviation of 19.4 seconds, and man-hours are overcalculated by approximately 18% when compared to the scheduled plan presented in Table 2. As captured by the images, the number of workers is not constant in the manufacturing of panels due to the variation of tasks and the process itself. Workers are often called to other workstations and some tasks require more workers than others. Hence, man-hours in Table 5 are calculated interactively according to the number of workers captured during the task duration period. Table 6 presents the duration of the manufacturing process for all recorded panels as determined by the proposed approach according to its construction task, which presents similar results compared with previous manual studies performed by Ritter et al. [32] at the floor panel station under study with a difference of 0.01% compared to the average duration per floor panel.

**Table 6.** Total duration of construction tasks for floor panels as determined by the proposed approach.

<table>
<thead>
<tr>
<th>Construction Task</th>
<th>Panel ID</th>
<th>Panel ID</th>
<th>Panel ID</th>
<th>Panel ID</th>
<th>Panel ID</th>
<th>Average Task Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle</td>
<td>16.67</td>
<td>8.42</td>
<td>8.17</td>
<td>11.50</td>
<td>38.42</td>
<td>2.00</td>
</tr>
<tr>
<td>Prep, Place and Secure Joists</td>
<td>16.58</td>
<td>6.92</td>
<td>12.50</td>
<td>25.42</td>
<td>23.17</td>
<td>7.33</td>
</tr>
<tr>
<td>Auto-gluing</td>
<td>8.25</td>
<td>3.50</td>
<td>5.17</td>
<td>5.17</td>
<td>2.17</td>
<td>2.08</td>
</tr>
<tr>
<td>Sheathing Placing</td>
<td>7.42</td>
<td>7.25</td>
<td>10.50</td>
<td>12.25</td>
<td>10.92</td>
<td>13.92</td>
</tr>
<tr>
<td>Auto-stapling and cutting</td>
<td>18.33</td>
<td>8.75</td>
<td>19.00</td>
<td>18.42</td>
<td>17.58</td>
<td>31.58</td>
</tr>
<tr>
<td>Finishing (Cleaning, Inspection and Lift Prep)</td>
<td>5.83</td>
<td>2.25</td>
<td>6.17</td>
<td>4.75</td>
<td>3.50</td>
<td>4.42</td>
</tr>
<tr>
<td>Crane Lifting</td>
<td>1.92</td>
<td>0.42</td>
<td>2.33</td>
<td>2.33</td>
<td>1.92</td>
<td>2.92</td>
</tr>
<tr>
<td><strong>Total Duration (min)</strong></td>
<td>75.00</td>
<td>37.50</td>
<td>63.83</td>
<td>79.83</td>
<td>97.67</td>
<td>64.25</td>
</tr>
<tr>
<td><strong>Task Efficiency (E)</strong></td>
<td>77.77%</td>
<td>77.55%</td>
<td>87.20%</td>
<td>85.60%</td>
<td>60.67%</td>
<td>96.89%</td>
</tr>
</tbody>
</table>

### 4.2 Discussion and limitations

As per Table 6, the overall task efficiency, \(E\), of floor panels is calculated as per Equation 6 and indicates the overall pace of production of the monitored task at the workstation under study.

\[
E = \frac{\sum_{i=1}^{n} d_x}{(\Delta t - b)}
\]  

where \((d_x)\) denotes the duration in which the VFSM is in all value-adding states to the task(s) (i.e. all states besides idle), \((\Delta t)\) is the elapsed time between all VSFM states and the end of its manufacturing process (end time on Crane Lifting), and \((b)\) is the duration of any programed break that may occur during the production of the panel. Programed break times (e.g., lunch break) must be removed from this indicator since it is a scheduled break, during which workers are not expected to work; therefore, the scheduled breaks cannot be considered as idle production time. As such, differentiation between scheduled breaks and idle time is recommended. By removing the break time and addressing the time spent on value and non-value-adding tasks, the proposed approach identifies how much time is wasted in current floor paneling operations and its impact in the total floor’s manufacturing duration. One of the advantages of using VFSM is its reconfigurability to adapt to new requirements: for example, Figure 10 shows the changes made to incorporate a new state, Scheduled Breaks, to the VFSM proposed in Figure 3.
where $b_n = \{[t_0, t_1], [t_2, t_3], ..., [t_{k-1}, t_k]\}$ is a list of tuples with the start and end times for each scheduled break and is a user-defined parameter, and (currentTime) is the actual time that needs to be added to the list of variables presented in Table 1. Although VFSMs can be easily modified to adapt to new production environments, they rely completely on a good modeling effort. VFSMs do not have the capacity of understanding variations to the current order of production or unexpected situations in the shop floor and are limited in trying to explain those situations properly as a great modeling effort is necessary to represent all the potential situations that can be observed (which may not be always possible).

The task efficiency of all floor panels recorded by the proposed approach is shown in Figure 11 and is compared to the average efficiency throughout the overall process. Panel 5 demonstrates the lowest efficiency caused due high idle times as indicated in Table 6. Despite being able to verify the original footage, the authors nor the proposed approach could not explain the reason for such an abnormal idle time in the production of Panel ID #5. Indeed, the footage provides vital information to the operations performed but does not provide the context required to fully identify abnormal situations. In this particular case, in which workers are not present, the idle time may have occurred due to lack of material, a scheduled safety training, or any other viable reason. Either way, more information is required to determine if the entire idle duration is, indeed, non-productive or part of it consist of scheduled breaks.
Based on the data shown in Table 6, Figure 12 shows the variation in the duration of all panels through a boxplot graph indicating a high variance in the Idle, Joists Placement, and Auto-stapling and Cutting states. The variation of task duration is important to evaluate and better predict the necessary resources and time required to manufacture each floor panel at the task level. The high variation in the idle state is expected due to the random nature of activities and events that causes delays in any productive task. However, root causes for these delays are not identified by the proposed approach, and may include, for example, machine breakdown, confusion caused by unclear construction drawings, lack of materials, etc. Moreover, idle times recorded by the proposed approach may be false positives in the case of scheduled events, such as safety training and meetings. These are limitations that will be further pursued in future research.

Figure 12. Task duration boxplot per task for floor panel manufacturing.

According to the analysis conducted by the authors, the placement of the joists and the auto-stapling process depends to a significant degree on the design features of the floor panels, such as the number of joists and the final shape of the panel. The identification of design features is not part of the proposed approach, but is essential for better prediction of construction task duration for planning purposes in future work. Indeed, the combination of the data recorded by the proposed approach and data from other sources, such as construction drawings and production schedules, affords a significant opportunity to better identify panel characteristics and address production’s ability to comply with deadlines. The proposed approach allows task durations to be detected and timed automatically with an average accuracy of 92% thus eliminating the need for experts to perform manual time studies and avoiding human errors. Video recordings of construction tasks are rich sources of information for the improvement of operations in offsite construction and should not be limited to productivity studies. Therefore, a holistic approach is recommended to address areas such as productivity, safety, and ergonomics by collecting relevant data from each area and recommending solutions to improve areas in an integrated manner.

5 Conclusions

The key to offsite construction monitoring is obtaining task progress information in a timely and continuous manner. In this study, taking floor panel manufacturing stations as an example, a vision-based approach for automated productivity evaluation of the progress of an offsite construction task is proposed. The proposed method can identify and locate key resources, using the Faster R-CNN machine learning technique, and then can model accordingly the sequence of tasks through virtual finite state machines. The neural network algorithm is trained to locate and identify workers and key equipment used for floor panel manufacturing using low-resolution footage from CCTV security cameras. The sequence modeling of task progress presented in this study as a virtual finite state machine addresses the current limitations for activity tracking of
machine vision approaches in low-resolution environments. By combining resource identification with task sequencing, the presented method can contextualize the construction activities for which resources are used. In addition, duration and man-hours used for each task can be calculated, providing managers with a clear and timely understanding of progress at the workstation.

The results presented demonstrate that the presented approach can calculate the duration and the man-hours required per task with an accuracy above 92%. The resulting automated analysis is comparable to similar manual time studies available in the literature and provides insights in terms of task duration and man-hours needed per panel at the workstation under study. The assumptions taken to simplify the proposed approach, e.g., use object detection instead of activity detection and assume that the workers detected in the station boundaries are actually working, do not affect in great measure the accuracy of the metrics calculated nor the task tracking.

On the basis of the results achieved so far, the authors will explore expanding the approach by including other data sources, i.e., RFID, and more characteristics of the activities performed at the station, such as maintenance operations or panel design information. Additionally, it could be interesting to estimate the working hours for a set of particular workers or crew and optimize the production schedule around their distribution across the manufacturing plant. Moreover, the authors intend to explore other aspects of production, such as safety and ergonomics, in order to make use of the available CCTV footage and enhance its application on these different areas.

Acknowledgements

The authors gratefully acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (File No. IRCPJ 419145-15).

References


