frontiers in Built Environment

REVIEW
published: 27 November 2020
doi: 10.3389/fbuil.2020.575738

Visual Analytics for Operation-Level Construction Monitoring and Documentation: State-of-the-Art Technologies, Research Challenges, and Future Directions

Jinwoo Kim[1,2]*

[1] Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, United States, [2] Institute of Construction and Environmental Engineering, Seoul National University, Seoul, South Korea

Operation-level vision-based monitoring and documentation has drawn significant attention from construction practitioners and researchers. To automate the operation-level monitoring of construction and built environments, there have been much effort to develop computer vision technologies. Despite their encouraging findings, it remains a major challenge to exploit technologies in real construction projects, implying that there are knowledge gaps in practice and theory. To fill such knowledge gaps, this study thoroughly reviews 119 papers on operation-level vision-based construction monitoring, published in mainstream construction informatics journals. Existing research papers can be categorized into three sequential technologies: (1) camera placement for operation-level construction monitoring, (2) single-camera-based construction monitoring and documentation, and (3) multi-camera-based onsite information integration and construction monitoring. For each technology, state-of-the-art algorithms, open challenges, and future directions are discussed.

Keywords: construction, site monitoring, operation-level, vision-based, state-of-the-art, computer vision, deep learning

OPEN ACCESS

Edited by:
Youngjib Ham,
Texas A&M University, United States

Reviewed by:
Yilong Han,
Tongji University, China
Jiayu Chen,
City University of Hong Kong, Hong Kong

****Correspondence:***
Jinwoo Kim
jinwooki@umich.edu;
jinwoo92@snu.ac.kr

Specialty section:
This article was submitted to Construction Management, a section of the journal Frontiers in Built Environment

Received: *24 June 2020*
Accepted: *24 August 2020*
Published: *27 November 2020*

Citation:
Kim J (2020) Visual Analytics for Operation-Level Construction Monitoring and Documentation: State-of-the-Art Technologies, Research Challenges, and Future Directions. Front. Built Environ. 6:575738. doi: 10.3389/fbuil.2020.575738

INTRODUCTION

Operation-level vision-based monitoring and documentation is vital to construction project managers because it enables to obtain useful information required for project management and control (Yang et al., 2015). By observing long-sequence video-streams collected by fixed cameras, managers can identify different types of activities performed by construction workers and equipment, and can measure their operational efficiency (e.g., direct work rate, hourly work amount) (Kim et al., 2018d, 2019d). For instance, if too many workers are waiting to commence concreting operations, managers can decide to allocate more concreting equipment (e.g., pump cars, mixer trucks) or reduce the number of onsite workers. Operation-level monitoring also allows managers to recognize unsafe behaviors of workers and equipment, such as lapses in wearing safety equipment, speed limit violations, and access to dangerous areas. Based on such operational information, project managers can perform safety training to prevent unsafe behaviors and minimize accident risks. In addition to construction phases, vision-based monitoring would

be also beneficial to operations and maintenance phases (Xu and Brownjohn, 2018). Accordingly, operation-level monitoring may be employed to complete construction projects successfully.

Given the importance of operation-level monitoring and documentation, project managers visit construction sites directly and gather onsite information manually. However, they often face difficulty in the continuous monitoring of dynamic and large-scale jobsites, and have thus installed closed-circuit television (CCTV) cameras at construction and built environments to perform vision-based monitoring remotely (Chi and Caldas, 2011; Kim, 2019). Since 2016, remote vision-based monitoring and documentation has received more attention as the Korean Government permitted construction companies to include the camera-installation expenses in their jobsite management budget (Korea Construction Technology Promotion Act, 2016). Nevertheless, because it still requires to process an excessive amount of CCTV videos manually, project managers reported several shortcomings associated with remote monitoring; it is labor-intensive, expensive, and time-consuming (Memarzadeh et al., 2013). Many researchers have thus developed computer vision technologies that can understand onsite images and extract useful information automatically. Previous research has shown the potential of computer vision technologies for diverse operation-level monitoring applications, such as productivity measurement (Bügler et al., 2017; Kim et al., 2017, 2019c) and safety analysis (Seo et al., 2015b).

Despite their encouraging findings, it remains a major challenge to exploit computer vision technologies in real-world construction sites owing to various theoretical and practical issues, implying that there are knowledge gaps in practice and theory. Hence, there have been a few studies to review computer vision research in construction and expand our understanding on state-of-the-art technologies (Seo et al., 2015a; Teizer, 2015; Yang et al., 2015; Ham et al., 2016; Xiao and Zhu, 2018; Fang et al., 2020b,c; Sherafat et al., 2020; Zhang et al., 2020b). However, because some of the review articles (Seo et al., 2015a; Teizer, 2015; Yang et al., 2015; Ham et al., 2016) were published before 2016, recent advances in deep learning and computer vision algorithms could not be covered. In addition, recent reviews focused only on limited monitoring purposes and/or technologies of operation-level monitoring, resulting in a lack of holistic understanding on state-of-the-art technologies. Specifically, Fang et al. (2020b; 2020c) and Zhang et al. (2020b) investigated existing studies in the perspective of safety monitoring. The reviews of Xiao and Zhu (2018) and Sherafat et al. (2020) were limited only to construction object tracking and action recognition technologies, respectively. Thus, knowledge gaps still remain unclear over the various purposes and technologies of operation-level monitoring and documentation. For example, in the context of productivity monitoring, there is a lack of understanding on how to install multiple cameras at complex jobsites, analyze the collected visual data, and integrate analysis results derived from a set of single cameras. To fill such knowledge gaps, this study comprehensively reviews 119 papers on vision-based construction monitoring and documentation, identifies open research challenges, and proposes potential future directions. State-of-the-art technologies, presented in top computer vision conferences, are also introduced to bridge knowledge gaps in the construction field. Through the holistic understanding, it would be possible to clarify what research topics should be investigated and suggest what state-of-the-art algorithms can be applied. The findings of this study can help practitioners to implement vision-based construction monitoring in real construction projects.

Figure 1 illustrates an overview of this review. This paper provides background knowledge on operational-level construction monitoring and documentation. It then explains a research methodology to retrieve, select, and review relevant journal articles. Next, previous research achievements, open challenges, and future directions are discussed extensively in the major three technologies of operation-level monitoring: (1) camera placement for operation-level construction monitoring, (2) single-camera-based construction monitoring and documentation, and (3) multi-camera-based onsite information integration and construction monitoring. Based on the results of in-depth review, the author further derives open research challenges and proposes future directions to address them, which does not fall in the above three technologies but have a significant impact when developing and implementing computer vision technologies. Finally, the author concludes the review and discusses the contributions of the research.

RESEARCH BACKGROUND

Operation-level vision-based monitoring and documentation is a process of collecting and analyzing construction site images continuously, thereby obtaining information required for operational performance analysis. It is generally performed to monitor onsite productivity and safety, which are two major performance indicators of construction operations. To achieve this goal, it is necessary to detect and track construction objects (e.g., workers, equipment, materials), recognize their individual actions, and understand object-to-object interactions as most construction operations are conducted by those three basic elements. For example, to carry out soil-loading operations, an excavator and a dump truck should be located near each other while performing specific individual actions (excavator: dumping, dump truck: stopping). To assess the proper use of hardhats, which is one of the major indicators of construction safety, project managers should be able to identify workers and hardhats, and to interpret their spatial interactions (e.g., distance, overlapping areas).

Within these contexts, project managers have installed multiple CCTV cameras at large-scale jobsites, observed the collected videos, and integrated operational information obtained from different cameras (see **Figure 2**). However, there is often some degree of difficulty when performing remote monitoring owing to cost and time limitations, and thus many researchers have endeavored to develop the operation-level vision-based monitoring technologies: (1) camera placement for operation-level construction monitoring, (2) single-camera-based construction monitoring and documentation, and (3) multi-camera-based onsite information integration and construction monitoring.

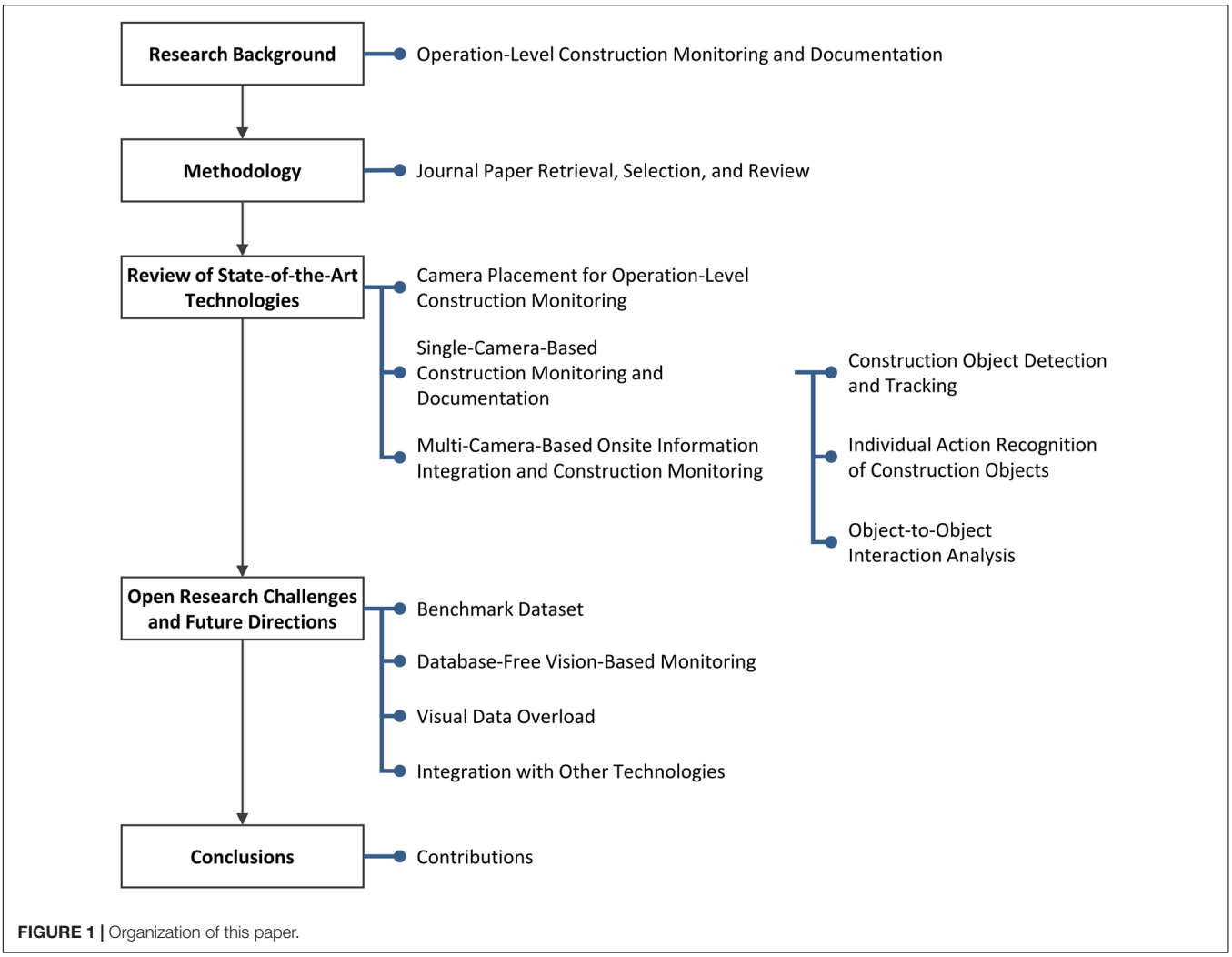

FIGURE 1 | Organization of this paper.

METHODOLOGY

Given this background, the author reviews existing works, identifies open research challenges, and proposes potential future directions. To do this, the present study leverages a content analysis approach for the systematic review of existing literature. It is a well-known approach for synthesizing the literature and deriving the objective findings, and its applicability has been extensively demonstrated in the areas of engineering and construction management (Yi and Chan, 2014; Mok et al., 2015; Liang et al., 2016; Li et al., 2018; Zhang et al., 2020b).

Figure 3 illustrates the research methodology consisting consists of three main processes: journal article collection, initial article review, and in-depth article review. To collect academic journal articles, an exhaustive search was carried out using the Web of Science, Scopus, and Google Scholar search engines. The scope of the publication search was limited to the period 2007/01/01 – 2020/07/31. This timeframe was selected because relevant publications have been published since 2007 (Zou and Kim, 2007). The search keywords can be grouped into two categories: construction- and technology-specific terms. The construction-specific terms include *construction site*, *construction operation*, *construction activity*, *site monitoring*, *construction management*, *workers*, *equipment*, *material*, *productivity*, and *safety*, while technology-specific terms include *computer vision*, *vision-based*, *image processing*, *image-based*, *machine learning*, *deep learning*, *surveillance*, *camera*, *video*, *detection*, *tracking*, and *recognition* so that the search can cover a broad range of related disciplines. These keywords enabled to collect a sufficient number of academic papers regarding operation-level construction monitoring and documentation. It is worth mentioning that, however, other terms related to project-level monitoring [e.g., drone, unmanned aerial vehicle (UAV)] (Kim et al., 2019d) were not included because this review focused on operation-level monitoring. The articles were retrieved from each search engine using the keywords, and the results were integrated through the removal of duplicate ones that have same digital object identifier. Based on the results, 176 articles were retrieved after which the initial review was performed as follows. The author also restricted the search to articles in mainstream journals, which were written in English as they are reputable and reliable sources (Zhong et al., 2019). Specifically, the well-known mainstream

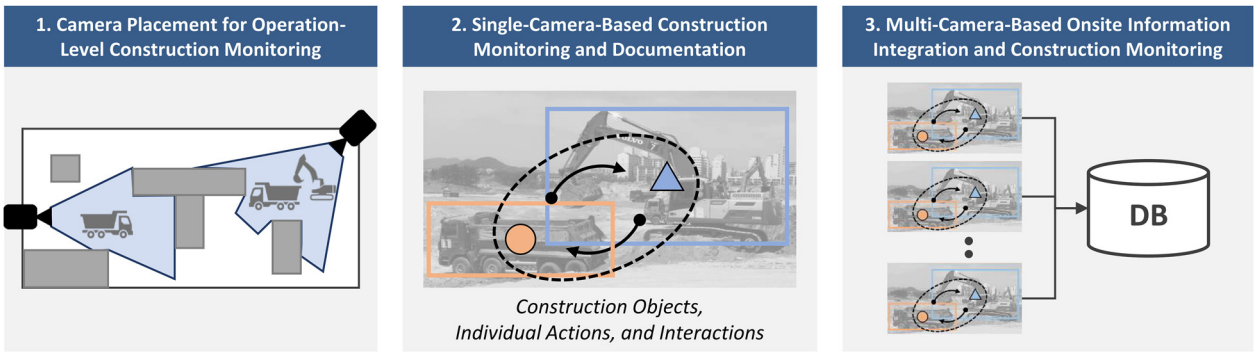

FIGURE 2 | Technologies of operation-level construction monitoring and documentation.

journals were chosen with reference to the findings of existing studies, which reviewed relevant research fields (Chen et al., 2018; Asadzadeh et al., 2020; Zhang et al., 2020b). They included *Frontiers in Built Environment*, *Automation in Construction*, *Advanced Engineering Informatics*, *Safety Science*, *Computer-Aided Civil and Infrastructure Engineering*, *Journal of Computing in Civil Engineering*, *Journal of Construction Engineering and Management*, *Journal of Management in Engineering*, *Journal of IT in Construction*, *Sensors*, *Journal of Civil Engineering and Management*, *Visualization in Engineering*, *KSCE Journal of Civil Engineering*, and *Canadian Journal of Civil Engineering*. Next, the publications that do not contain the aforementioned keywords in their titles or abstracts were filtered out, and less relevant and irrelevant papers were also screened out after a brief visual assessment of the content of articles. In this stage, two criteria were considered: (1) focus on operation-level monitoring (i.e., onsite productivity and safety analysis) and (2) focus on computer vision technologies or technologies integrated with computer vision. For example, journal papers that heavily focused on camera-equipped-drone-based project-level monitoring were filtered out; however, some of them are introduced to discuss an opportunity to integrate digital cameras and drones in a later section. Eventually, 119 journal papers in relevant to the monitoring of construction and built environments were considered in the in-depth article review stage. The journal papers were classified into each technology of vision-based monitoring and documentation as shown in **Figure 2**, and the open research challenges and future directions were discussed via content analysis from a theoretical and practical point of view.

REVIEW OF STATE-OF-THE-ART TECHNOLOGIES

Camera Placement for Operation-Level Construction Monitoring

This technology aims to install multiple cameras at different physical locations and collect appropriate video data from construction and built environments. It can be defined as a problem that determines the proper number, types, locations, and orientations of cameras required to sufficiently monitor a large-scale jobsite. According to one recent study (Kim et al., 2019d), project managers generally carry out the camera placement based on their knowledge or experiences rather than systematic guidelines. Their decisions occasionally come with the acceptable performance of vision-based monitoring (in terms of camera coverage or total costs), but they often experience difficulty in determining appropriate camera configurations and recording video-streams owing to the complex and dynamic natures of a given jobsite (e.g., power supply, data transmissibility, occlusion effects). For these reasons, researchers have attempted to find the optimal camera placement for vision-based construction monitoring.

As a first step, several studies investigated how to visualize and quantify visible coverages of installed cameras (**Figure 4**). Chen et al. (2013) developed a visible coverage visualization technique for measuring the performance of camera networks in public building spaces. Albahri and Hammad (2017a) also proposed a method that calculates the coverage of cameras installed in indoor buildings using Building Information Modeling (BIM). Building upon the coverage visualization and quantification, researchers made efforts to optimize the camera placement through the integration with metaheuristic algorithms. In one study (Albahri and Hammad, 2017b), the BIM-based coverage calculation method and a genetic algorithm were combined to find optimal configurations of camera networks in indoor buildings. Particularly, BIM played a key role in automatically deriving installation conditions such as geometrical constraints (e.g., ceiling) and operational conditions (e.g., vibrations produced by facility equipment). As the existing studies focused on optimizing the camera placement in BIM-available indoor buildings, other researchers studied how to place digital cameras in outdoor construction environments. For example, Yang et al. (2018) built a two-dimensional (2D) spatial model of a given jobsite and optimized the multi-camera placement using a genetic algorithm. Zhang et al. (2019) enhanced the jobsite modeling techniques to generate a three-dimensional (3D) virtual space, which is more effective in evaluating the visible coverage of camera networks in real-world construction environments. To further improve the practical usefulness of previous studies,

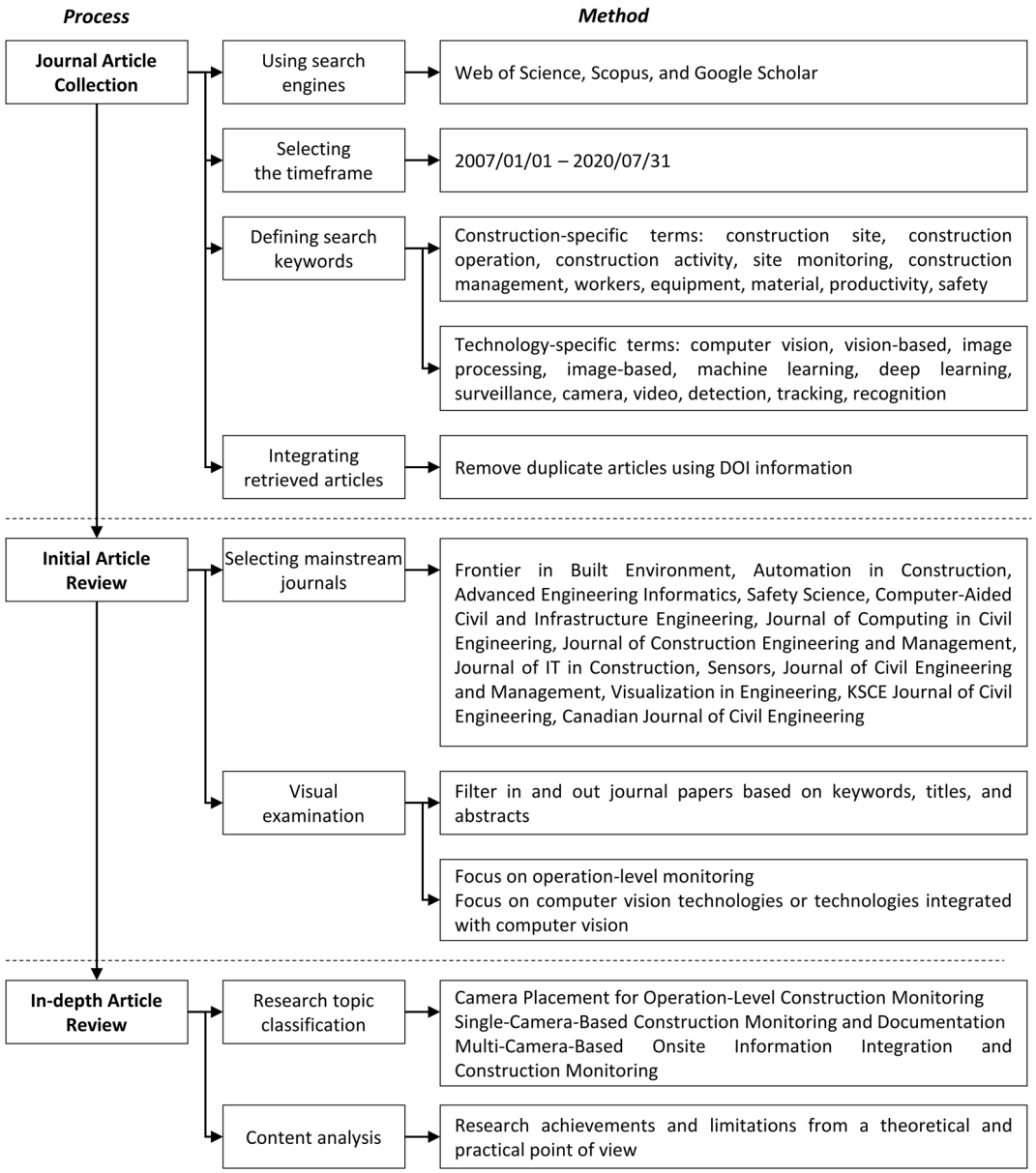

FIGURE 3 | Research methodology.

Kim et al. (2019c) discovered major practical characteristics of construction sites and camera networks to be considered when installing cameras on jobsites (e.g., power supply, data transmission). The characteristics were integrated with the previous optimization method, thereby producing network alternatives that are more applicable to actual construction sites.

Earlier studies showed encouraging results in optimizing the multi-camera placement and obtaining video data for operation-level vision-based monitoring. However, because only a few studies have been conducted in this research area, major research challenges remain unaddressed. One of the most urgent issues is that the dynamic and complex natures of construction and built environments has not been fully considered. Although one study incorporated various practical constraints of construction sites, such as power accessibility and data transmissibility (Kim et al., 2019d), most optimization techniques are based on an underlying assumption: construction sites are static, which means that jobsite components (e.g., structures, management offices, work zones) do not vary over time. Thus, existing camera networks for given jobsites are frequently unable to obtain appropriate video data (e.g., objects-of-interest are invisible from cameras), and further, they may need to be relocated as construction operations proceed, resulting in additional costs and monitoring difficulty. To address this issue, it is necessary to find the optimal camera configurations that are robust to time-dynamic conditions of jobsites. For instance, future studies can feed a four-dimensional (4D) BIM

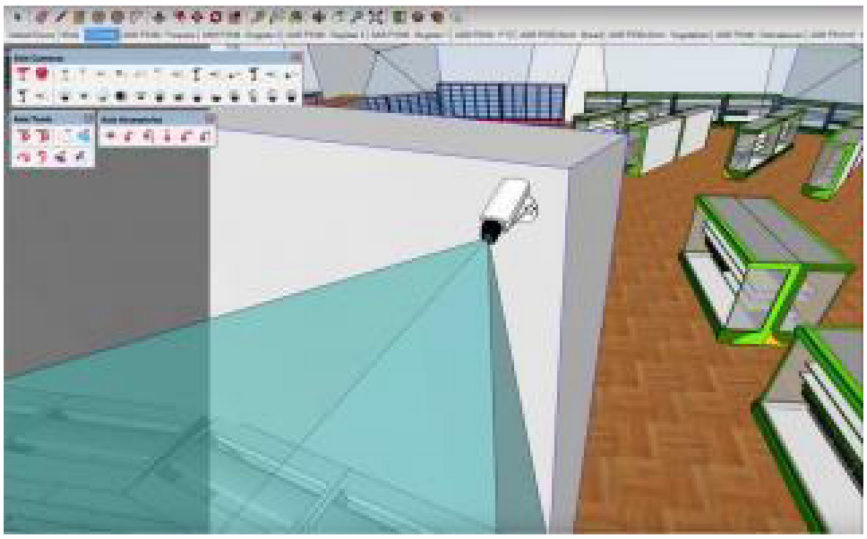

FIGURE 4 | Example of camera coverage in indoor buildings (AXIS Coverage Shape Software).

model (*x, y, z, t*) into optimization processes to consider jobsite constraints and conditions varying over time. Otherwise, if possible to relocate camera networks several times (e.g., three times during a project duration), previous approaches can be applied repeatedly for a few important stages (e.g., earthmoving, framing, finishing).

Next, since earlier works developed camera placement methods for limited monitoring purposes and project types, it is still burdensome to determine a generalized camera placement framework and optimal configurations. In this regard, the approach proposed in the previous research (Kim et al., 2019d) can become a baseline for deriving and structuralizing diverse characteristics of construction sites. They categorized the characteristics of building construction sites into visual monitoring determinants, influencing factors, and camera placement conditions (i.e., camera network and jobsite conditions) based on the result of in-depth interviews with 12 construction experts. Building on this taxonomy, future research can discover more various considerations by performing expert interviews and case studies from diverse projects types. Once the taxonomy is built, it would be beneficial to develop camera placement frameworks that are applicable to various monitoring purposes and project types.

Finally, there were no research efforts to optimize the camera placement from the perspective of maximizing the performance of visual analytics. Existing studies mainly focused on maximizing the visible coverage of camera networks and minimizing installation costs. However, for automated vision-based monitoring, it may be more significant to find the optimal configuration that can well-extract the operational information, rather than only maximizing the visible coverage. In this sense, researchers can put the first attempt to optimize the camera placement that maximizes the performance of construction object detection, which is a prerequisite step for operation-level vision-based monitoring. They can utilize a 4D jobsite simulation model, developed in many previous studies (Boton, 2018; Swallow and Zulu, 2019; Wang et al., 2019a; Choi et al., 2020), to place cameras and evaluate the monitoring performances (e.g., object detectability) in virtual spaces.

Single-Camera-Based Construction Monitoring and Documentation

After digital cameras are installed at jobsites, it is required to analyze the visual data collected from each single-camera and extract onsite information. As described above, previous research focused on capturing operational information about construction objects, their individual actions, and object-to-object interactions from jobsite images.

Construction Object Detection and Tracking

A large number of research papers were found in the area of construction object detection and tracking (**Table 1**). This may be because it is fundamental and essential to obtain sequential locations of construction objects for operation-level monitoring (**Figure 5**). For example, Chi et al. (2009) applied spatial modeling and image matching techniques to identify and track construction objects in laboratory environments. Park and Brilakis (2012) used a background subtraction method to localize construction workers from actual jobsite videos, and they achieved the detection rate of 99.0%. Several studies assessed the performance of visual tracking algorithms (e.g., mean-shift tracking, Bayesian contour tracking, active contour tracking) when localizing a single construction object (Teizer and Vela, 2009; Brilakis et al., 2011; Park et al., 2011). Chi and Caldas (2011) trained machine learning classifiers, such as Bayes models and neural networks, and compared their performances for object identification. Azar and McCabe (2012a) integrated a support vector machine (SVM) with histogram-of-oriented-gradient (HOG) features to recognize dump trucks from onsite images, Memarzadeh et al. (2013) presented a similar approach

TABLE 1 | Summary of existing object detection and tracking algorithms.

Algorithms	Objects-of-interest	Dataset (image frames)		Performance (accuracy or error distance)	Literature
		Training	Test		
Spatial modeling Image matching	Workers	N/A	N/A	85.0%	Chi et al. (2009)
Background subtraction	Workers	2,200	500	99.0%	Park and Brilakis (2012)
Mean-shift tracking Bayesian contour tracking Active contour tracking Graph-cut tracking	Workers	N/A	3,730	2.12 pixel	Teizer and Vela (2009)
Contour-based tracking Kernel-based tracking	Workers and equipment	33	35	82.3%	Brilakis et al. (2011)
Contour-based tracking Kernel-based tracking Point-based tracking	Workers and equipment	N/A	2,037	17.6 pixel	Park et al. (2011)
Bayes classifier Neural network classifier	Workers and equipment	750	1,282	96.7%	Chi and Caldas (2011)
HOG SVM	Dump trucks	800	380	86.8%	Azar and McCabe (2012a)
	Excavators	770	253	95.2%	Azar and McCabe (2012b)
HOG + color SVM	Workers and equipment	12,148	5,359	88.6%	Memarzadeh et al. (2013)
Particle filtering	Workers and equipment	N/A	3,080	11.1 pixel	Zhu et al. (2016)
Functional integration of detection and tracking	Workers	9,664	4,793	98.2%	Park and Brilakis (2016)
	Workers and equipment	44,597	8,066	86.5%	Zhu et al. (2017)
Functional integration of detection and tracking Online learning	Equipment	N/A	64,968	86.4%	Kim and Chi (2017)
Adaptive object model	Workers	688	2,079	72.2%	Konstantinou et al. (2019)
Faster R-CNN	Equipment	2336	584	96.3%	Kim et al. (2018b)
	Workers and equipment	8,500	1,500	93.0%	Fang et al. (2018e)
	License plates of dump truck	6,851	387	95.8%	Kim et al. (2019b)
Faster R-CNN Residual network	Workers	Public dataset	2,241	94.3%	Son et al. (2019)
SSD	Equipment	2,617	654	91.2%	Arabi et al. (2020)
	Equipment	216	24	98.8%	Guo et al. (2020)
Mask R-CNN Residual network	Workers, equipment, and materials	2,000 Public dataset	800	0.2 m	Fang et al. (2020a)
Mask R-CNN Gradient-based algorithm	Workers	Public dataset	Public dataset	81.8%	Angah and Chen (2020)
LSTM	Workers and equipment	152,000	58,000	1.11 pixel	Tang et al. (2020a)

for classifying workers and equipment simultaneously; they increased the detection accuracy from 86.8% to 88.6% while diversifying types of detectable objects. The SVM-HOG approach was also used to detect deformable parts of earthmoving excavators, e.g., buckets, arms, main bodies, in another study (Azar and McCabe, 2012b).With additional considerations of deformable parts, the performance of the SVM-HOG approach was improved up to 95.2%. To continuously track target objects in complex construction environments, Zhu et al. (2016) tested the robustness of particle filter algorithms to occlusion effects, which are a major cause of visual analysis failures, and the error rates were reported as 11.1 pixels. Park and Brilakis (2016) and Zhu et al. (2017) also made efforts to handle the occlusion issues through the functional integration of detection and tracking. Such detection-tracking integrated frameworks were further enhanced by Kim and Chi (2017). They appended online learning techniques that generate object-specific training data in real-time, enabling to train more powerful object detectors and track target objects better. Their method was able to achieve 86.4% accuracy without human-labeled training data. Similar to the concept of online learning, Konstantinou et al. (2019) proposed an adaptive model to track construction workers continuously.

With the remarkable advances in deep learning algorithms, researchers have investigated deep neural networks (DNNs) for construction object detection and tracking. Kim et al. (2018b) designed a region-based convolutional neural network (R-CNN) for the detection of various types of construction equipment (accuracy: 96.3%). As most objects were located at near the center in the testing images, Fang et al. (2018e) further applied a Faster R-CNN model for the real-world construction scenes and showed the detection rate of 93.0%. Son et al. (2019) also improved the previous R-CNN model by combining residual neural networks to recognize construction workers under varying postures and viewpoints (accuracy: 94.3%). Single shot detector

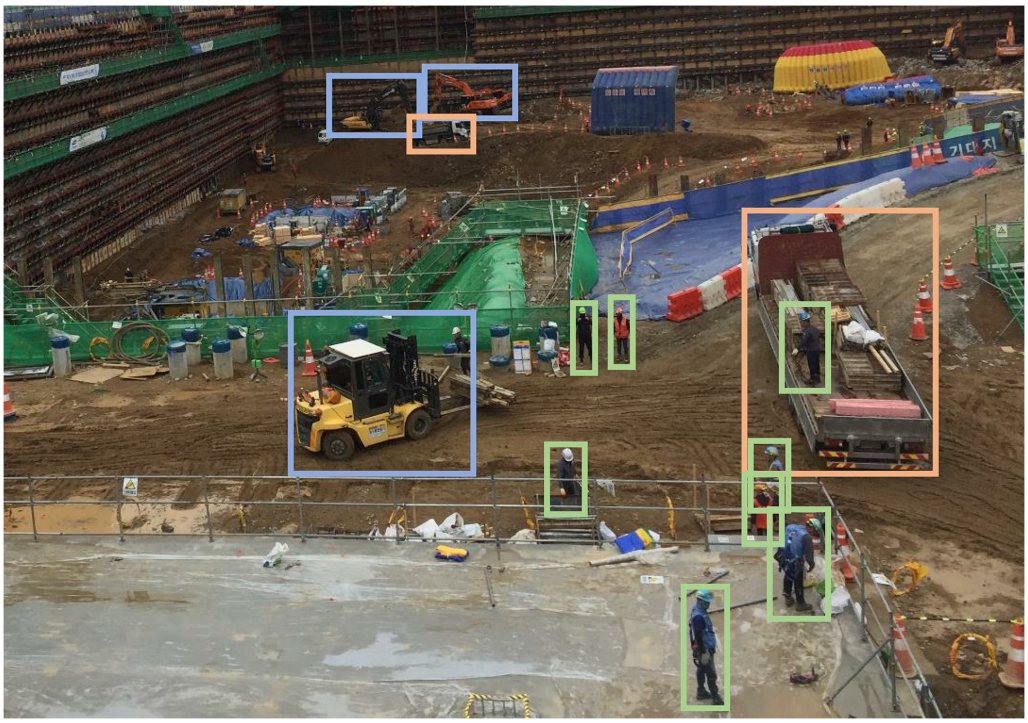

FIGURE 5 | Examples of construction object detection.

(SSD), which is a lighter CNN model but which has a comparable performance to the R-CNN, was also tested in another study (Arabi et al., 2020); in their experiments, the SSD model showed 91.2% detection rate. Other researchers further examined the applicability of DNNs to detailed visual analyses. Kim et al. (2019b) used a Faster R-CNN model to detect license plates of dump trucks and obtain their site-access data, and Guo et al. (2020) built an SSD-based system for detecting construction entities and capturing their physical orientations. The CNN-based method presented in another study (Fang et al., 2020a) localized construction workers and equipment at pixel-level and the average error rate was reported as 0.2 m. Similarly, Angah and Chen (2020) integrated a Mask R-CNN model with a gradient location prediction algorithm to detect and track construction workers with the accuracy of 91.2%. Tang et al. (2020a) also predicted motion trajectories of construction objects using LSTM encoder-decoder model. In their experiments, the average error rate was reported as 1.11 pixels.

Despite the successful achievements, significant research opportunities remain to be addressed in the field of construction object tracking. One of the most challenging issues is to track multiple objects over the long term from complex construction images. As there exist a large number of construction objects that move dynamically, it is common for target objects to be occluded by others, causing to disappear and reappear from the camera's field-of-view (FOV), resulting in frequent tracking failures. For this issue, it is recommended to investigate how to re-track construction objects when they reappear in a camera's FOV after the occlusions and/or disappearances. One attractive solution is to adopt the Markov Decision Process (MDP) approach proposed by one study (Xiang et al., 2015). The MDP approach considers re-tracking as a problem of undertaking one of the three transition actions when a target object has the state of *lost*: a *lost* target can be *lost* again, *tracked* if it reappears, and *inactive* if it disappears forever. Here, the state of *lost* represents when a target object was not detected owing to occlusions and/or disappearances. Given this formulation, researchers need to carefully combine decision-making criteria with deep decision networks. For example, most state-of-the-art algorithms leverage deep visual features of target objects (extracted by CNN) under an assumption: a target object may have a visual similarity before and after the occlusions and disappearances (Guo and Cheung, 2018; Chen Y. et al., 2020). Spatial information is also an important source for re-tracking a lost object; a lost object may reappear in near areas where it was occluded or disappeared (Shen et al., 2018). However, since construction objects and surrounding environments changes over time, it would be also effective to consider temporal dynamics, such as sequential patterns of object movements (e.g., locations and speeds over time). In the computer vision domain, Milan et al. (2017) employed a recurrent neural network (RNN) to predict possible regions-of-interest of a lost object in the future image frames. In addition, recent computer vision studies have also emphasized the importance of interaction features for object re-tracking (Sadeghian et al., 2017; Gupta et al., 2018). The interaction features represent a situation where an object's movements (e.g., speed, directions) may be affected by other objects in crowded environments. This phenomenon may be

also observed in construction scenes. For instance, a worker would bypass nearby heavy equipment and/or hazardous areas (e.g., holes on worksites). Such interaction features can be effectively modeled with graph neural networks (GNNs), which are an emerging deep learning architecture. The GNN forms target objects and their interactions using a set of graph nodes and edges, and learns object-to-object relationships, thereby well-encoding interaction effects among different objects (Qi et al., 2018). By properly selecting purpose-specific features and decision networks, it would be possible to effectively represent the dynamic movements of construction objects and re-track them if lost.

Another major challenge is that it remains unclear exactly what technical problems should be solved for construction object tracking. This is because construction researchers have tested tracking algorithms under different analytics conditions (Xiao and Zhu, 2018). For the first step, it is vital to build and share benchmark datasets, as argued in many existing studies (Seo et al., 2015a; Kim et al., 2018b). In addition to the benchmark datasets, it may be also helpful to develop a simple baseline model, which does not require any tracking-specific training and complex optimization tasks, and comprehensively analyze its performance and failure cases. Interestingly, recent studies in the computer vision domain reported that a very straightforward algorithm (i.e., intersection-over-union-based tracker) has a comparable performance with deep learning-based cutting-edge tracking algorithms (Bochinski et al., 2018; Bergmann et al., 2019). These findings motivated researchers in the field of deep learning and computer vision to revisit the existing tracking paradigm, which is based on complex DNNs. Based on the experimental results, the authors analyzed in-depth several failure cases and derived the primary failure causes, providing guidelines for future research. It would be also valuable if similar approaches are applied to construction sites, and specific problems can be identified and addressed by future studies.

Individual Action Recognition of Construction Objects

The second step of single-camera-based construction monitoring and documentation is to recognize the individual actions of construction objects using their detection and tracking results, i.e., spatio-temporal trajectories. The information about individual actions of onsite workers (e.g., working, idling) can be used for operational performance analysis. In this area, previous methods mainly focused on two different types of monitoring purposes, i.e., productivity and safety analysis (**Table 2**). For instance, Zou and Kim (2007) proposed a method that analyzes hue, saturation, and value color spaces to classify whether an excavator is working or idling. Golparvar-Fard et al. (2013) adopted a concept of bags-of-visual features to categorize activity types of earthmoving excavators and dump trucks. Their method could recognize the activity types with the performance of 91.2%. In one study (Soltani et al., 2017), deformable parts of excavators and their detailed postures were detected with 84.0% accuracy. Gong and Caldas (2010) trained a machine learning classifier to identify tower cranes' buckets and to track their movements for the productivity analysis of concrete placement operations. This method was improved and applied to other types of operations, such as earthmoving, slab pouring, hoisting, and scaffold installation (Gong and Caldas, 2011). The productivity of tower cranes was also analyzed in another study (Yang et al., 2014). Their method inferred the activity types of tower cranes using site-layout information and operation cycles. Vision-based approaches were also developed for the productivity analysis of construction workers. Gong et al. (2011) combined bags-of-visual features and Bayesian networks to classify the actions of construction workers, such as traveling, transporting, and aligning. Yang et al. (2016) tested the performance of different features of HOG, histogram-of-optical-flows, and motion-boundary-histograms. The performance of their method was limited to 58.5% for more than 10 different operations, but they showed the possibility of diversifying the types of detectable activities. In recent years, similar to object detection and tracking, further studies have been performed to apply deep learning algorithms. Chen C. et al. (2020) designed a rule-based reasoning algorithm to interpret the working states of excavators from the CNN detection results, and their performance was 87.6%. Roberts and Golparvar-Fard (2019) also employed a hidden Markov model to analyze the tracking results of earthmoving equipment (i.e., bounding boxes over time) and to identify their activity types (accuracy: 85.1%). Kim and Chi (2019) improved the performance of excavator action recognition by integrating CNN and double-layer LSTM (90.0%). In other studies (Liang et al., 2019; Luo et al., 2020b), the detailed skeletons and postures of earthmoving excavators were detectable by CNN-based models. Luo et al. (2018c) proposed a two color-temporal stream CNN model that recognizes workers' actions, and the model was further improved in later studies. Luo et al. (2018a) added one more gray-scale color stream to improve the model performance (80.5% → 84.0%), and Luo et al. (2019) appended Bayesian nonparametric learning to capture workers' activities in far-field surveillance videos. In the research (Roberts et al., 2020), the CNN-based detection results were used to track multiple workers and estimate their detailed postures and joint angles. They could achieve the performance of 82.6% while providing more detailed information (i.e., postures, joint angles) compared to other action recognition research.

Regarding safety analysis, researchers mainly focused on monitoring unsafe behaviors of construction workers who are the major victims of accidents. Han and Lee (2013) identified safety-specific physical movements and presented a vision-based method that captures the workers' motions, and these results demonstrated the applicability of computer vision technologies for construction safety monitoring. Han et al. (2013a) conducted an empirical assessment of depth cameras for workers' motion capture and recognition, and they used the motion information to detect unsafe actions of workers when performing a ladder climbing activity (Han et al., 2013b). The research team also tested diverse types of motion features, such as joint angles, position vectors, and movement directions, to detect workers' unsafe actions in a later study (Han et al., 2014). Thereby, they could increase the performance of action recognition to 94.6%. Yan et al. (2017) further improved the posture recognition method through the application of view-invariant

TABLE 2 | Summary of existing action recognition algorithms in construction.

Algorithms	Objects-of-interest	Dataset (image frames)		Performance (accuracy or error distance)	Literature
		Training	Test		
Productivity monitoring					
Color Rule-based reasoning	Excavators	300	1,080	99.8%	Zou and Kim (2007)
Gaussian background subtraction Rule-based reasoning	Tower cranes	N/A	7,287	87.5%	Yang et al. (2014)
Boosted cascade Rule-based reasoning	Tower cranes and buckets	4,260	234,000	97.5%	Gong and Caldas (2010)
	Workers and equipment	12,060	72,000	86.1%	Gong et al. (2011)
Bags-of-HOG features SVM	Excavators and dump trucks	N/A	N/A	91.2%	Golparvar-Fard et al. (2013)
	Workers	1,570	720	58.5%	Yang et al. (2016)
HOG SVM	Excavators	30,600	557,625	84.0%	Soltani et al. (2017)
Faster R-CNN Deep SORT tracker Rule-based reasoning	Excavators and dump trucks	19,260	5,280	87.6%	Chen C. et al. (2020)
Retina networks Hidden Markov model	Excavators and dump trucks	29,518	59,037	85.1%	Roberts and Golparvar-Fard (2019)
CNN-LSTM	Excavators	40,000	32,365	90.0%	Kim and Chi (2019)
Hourglass networks Cascaded pyramid networks	Excavators	3,000	500	144.6 mm	Liang et al. (2019)
	Excavators	5,124	1,281	90.0%	Luo et al. (2020b)
Multi-stream CNN	Workers	63,300	31,650	80.5%	Luo et al. (2018c)
	Workers	47,136	20,202	85.0%	Luo et al. (2018a)
Temporal segment networks Bayesian nonparametric learning	Workers	27,000	13,500	84.0%	Luo et al. (2019)
YOLOv3 Single-person pose estimators	Workers	26,504	13,252	82.6%	Roberts et al. (2020)
Safety monitoring					
HOG Mixture-of-parts model	Workers	220	5,405	88.0%	Han and Lee (2013)
Kinect motion capture systems	Workers	N/A	3,136	100%	Han et al. (2013a)
	Workers	1,385	8,310	90.9%	Han et al. (2013b)
SVM	Workers	1,385	8,310	94.6%	Han et al. (2014)
DNN Decision tree k-Nearest Neighbor	Workers	75,600	32,400	95.0%	Yan et al. (2017)
OpenSim 3D SSPP	Workers	N/A	3,600	95.5%	Seo et al. (2015b)
Point tracking-based model	Workers	N/A	N/A	3.8 cm	Liu et al. (2016)
CNN	Workers	8,000 Public dataset	180,000	83.4%	Fang et al. (2018c)
	Workers	Public dataset	15,779	94.9%	Zhang et al. (2018b)
	Workers	N/A	1,000	96.7%	Chu et al. (2020)
CNN-LSTM	Workers	56,250	18,750	92.0%	Ding et al. (2018)
Hourglass networks	Workers	Public dataset	3,878	3.9 cm	Yu et al. (2019b)
	Workers	Public dataset	3,878	85.0%	Yu et al. (2019a)

features, resulting in a slight performance gain (95.0%). In another study (Seo et al., 2015b), the detected motions and postures were utilized to perform workers' biomechanical and musculoskeletal risk assessment; they fed the motion data into the existing biomechanical analysis tool, i.e., 3D static strength prediction program. Liu et al. (2016) estimated worker's 3D skeleton from stereo video cameras for ergonomic analysis. In line with productivity monitoring, deep learning approaches also shown promising results in construction safety analysis. Fang et al. (2018c) utilized a CNN-based face identification model to confirm whether workers are performing a non-certified operation (83.4%), and Ding et al. (2018) proposed a hybrid deep learning model composed of CNN and LSTM to detect unsafe behaviors of construction workers when climbing a ladder (92.0%). To interpret workers' unsafe motions in detail, Zhang et al. (2018b) extracted their 3D postures and body joints from a single ordinary camera with 94.9% accuracy. Such fundamental posture data were further analyzed for physical fatigue assessment (Yu et al., 2019b), biomechanical workload estimation (Yu et al., 2019a), and ergonomic posture analysis (Chu et al., 2020).

Although much research focused on vision-based action recognition in construction, there are several open challenges to be addressed. First, existing approaches were designed only for limited types of construction operations and objects. As for productivity monitoring, most studies analyzed the

operational efficiency of earthmoving excavators and dump trucks, and only a few studies evaluated the productivity of framing operations. Hence, it remains unclear what types of construction objects and individual actions need to be recognized for different operations, e.g., foundation and finishing works. Future researchers should develop a construction-specific dictionary that categorizes objects-of-interest as well as their meaningful behaviors in terms of productivity monitoring. For instance, to measure the cycle time and productivity of curtain-wall installation, which is a widely performed operation in recent building construction projects, it is recommended to track mobile/tower cranes, curtain-walls, and workers and to identify their individual actions (e.g., lifting). Regarding the safety monitoring, researchers primarily focused on construction workers and their unsafe postures/behaviors when performing specific operations, such as ladder climbing and masonry works. Accordingly, it is proposed that computer vision technologies be applied to a wider variety of construction operations for safety monitoring purposes. In particular, according to the interviews with construction safety experts, there is a great interest in monitoring dangerous behaviors, such as smoking, smombing (i.e., walking while on the phone), and unconsciousness.

Next, there is a limitation to develop high-performance vision-based algorithms that are applicable to actual construction sites. Because construction objects can conduct different actions while showing similar visual characteristics, it is difficult to classify those actions using only visual signals. For example, both "swinging" and "hauling" excavators rotate their deformable parts (e.g., main body, arm, bucket), and "sitting" workers can perform "steel-framing" or "resting." Operational contexts can play a key role to address this issue. In one study (Kim and Chi, 2019), they considered sequential working patterns of visual features and operations cycles to recognize various types of excavator actions (**Figure 6**), thereby improving the recognition performance by 17.6%. Another major cause of performance deviations is the low image resolution. Construction objects are often observed with a low resolution in a camera's FOV, resulting in a lack of visual information and analysis difficulty. Such intrinsic shortcomings can be effectively overcome by using pan-tilt-zoom (PTZ) cameras. If using PTZ cameras, it would be possible to select an object-of-interest from an entire-jobsite view, PTZ the object, and obtain a sufficient image resolution for visual analysis. In this way, individual actions of construction objects can be well-recognized, even with existing algorithms.

Object-to-Object Interaction Analysis

Researchers also investigated how to interpret the object-to-object interactions and relationships from complex construction scenes (**Table 3**). Azar et al. (2013) first developed a system that analyzes excavator-to-dump truck interactions and classifies soil-loading activities, and Bügler et al. (2017) adopted the interaction analysis system to estimate earthmoving productivity. Kim et al. (2018c) further improved the previous method by considering the action consistency of equipment. The methods developed in other studies (Bügler et al., 2017; Kim et al., 2018c) could achieve the performance over 90% without requiring a pretrained model. As these studies focused only on the interactive operations of earthmoving equipment, Luo et al. (2018b) developed a relevance network that infers the likelihood of interactions among workers, equipment, and materials during the framing operations. Their method was able to detect various activities of different resources, but the performance remained in the vicinity of 67.3%. This method was further enhanced to identify working groups, i.e., which objects are working together for a certain operation in a later study (Luo et al., 2020c). Cai et al. (2019) also designed a two-step LSTM model to achieve working group identification and activity classification. Interaction analysis has been also studied for safety monitoring purposes. The method presented in a study (Chi and Caldas, 2012) assessed the probability of struck-by-object accidents based on object-to-object spatial interactions (e.g., proximity). Kim et al. (2016) also monitored struck-by-objects using the spatial relationships of construction workers and equipment; the performance was slightly decreased by 2.9% compared to the results of the previous study (Chi and Caldas, 2012), but their fuzzy inference method enabled construction workers to be aware of unsafe conditions rapidly. Zhang et al. (2020a) customized the spatial analysis method to evaluate their collision risks between excavators and workers. Hu et al. (2020) used more detailed spatial features of excavators and workers (i.e., skeletons) for collision risk assessment; the approach achieved 91.6% accuracy while providing spatial information more precisely. In another study (Fang et al., 2019), their method detected construction workers and structural supports, and examined their spatial relationships to prevent falls from heights. As wearing personal-protective-equipment (PPE) is a major indicator that indicates the levels of workers' safety, other researchers explored the spatial interactions between workers and PPEs. Park et al. (2015) presented a vision-based system that localizes both workers and hardhats, interpreted their spatial interactions (e.g., overlapping areas), and determines non-hardhat-wearing states. The non-hardhat-wearing detection system was further enhanced to be applicable to far-field surveillance images (Fang et al., 2018b) and indoor dynamic environments (Mneymneh et al., 2019). Similar systems were also developed for other types of PPEs, such as safety vests (Nath et al., 2020), harnesses (Fang et al., 2018d), anchorages (Fang et al., 2018a), and eye/face/hand protection (Tang et al., 2020b). Owing to the intrinsic limitation of 2D image analysis, there have been several studies to capture the 3D spatial relationships of construction objects. Yan et al. (2019b) proposed a safety assessment method that detects construction workers and calculates 3D spatial crowdedness using a single camera. Luo et al. (2020a) presented a 3D proximity estimation technique for preventing worker-to-excavator collisions. Yan et al. (2020) enhanced the single-camera-based system to predict 3D spatial relationships of various construction objects, e.g., workers, excavators, dump trucks, and mixer trucks.

Despite the promising results, it is still challenging to understand various interactions and relationships among construction objects from complex onsite images. One of the most crucial issues is that existing studies focused on monitoring limited types of construction operations for specific purposes. Productivity monitoring research was limited mainly to earthmoving and framing operations, and vision-based safety

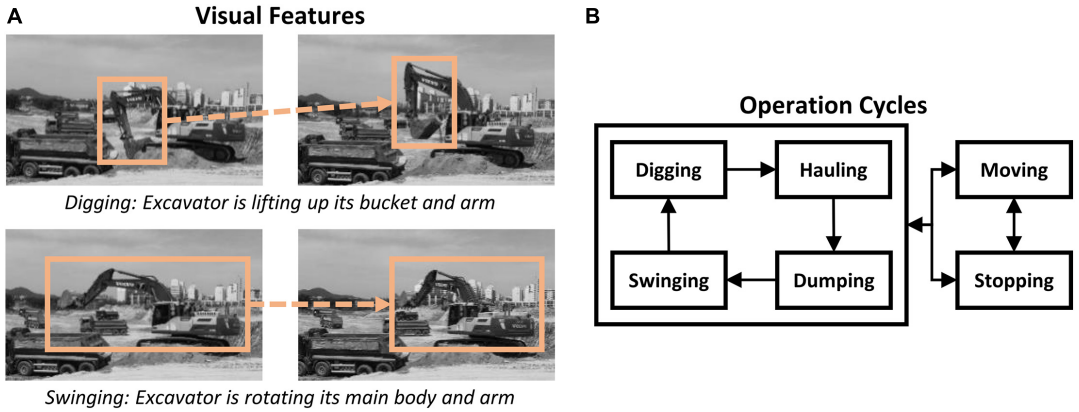

FIGURE 6 | Examples for the sequential working patterns of earthmoving excavators: **(A)** Sequential patterns of visual features. **(B)** Sequential patterns of operation cycles. Reprinted from Kim and Chi (2019) This article was published in *Automation in Construction*, Kim, J. and Chi, S. (2019) Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles. vol. 104, pp. 255–264, Copyright Elsevier (2019).

TABLE 3 | Summary of existing interaction analysis algorithms in construction.

Algorithms	Objects-of-interest	Dataset (image frames)		Performance (accuracy or error distance)	Literature
		Training	Test		
Productivity monitoring					
SVM	Excavators and dump trucks	1,342	180,000	95.0%	Azar et al. (2013)
Rule-based reasoning	Excavators and dump trucks	N/A	144,000	90.0%	Bügler et al. (2017)
	Excavators and dump trucks	N/A	11,513	91.2%	Kim et al. (2018c)
Relevance networks	Workers, equipment, and materials	6,232	1,558	67.3%	Luo et al. (2018b)
Conditional random field	Workers, equipment, and materials	396,536	188,160	98.7%	Luo et al. (2020c)
LSTM	Workers and equipment	1,710	428	95.0%	Cai et al. (2019)
Safety monitoring					
Rule-based reasoning	Workers and equipment	N/A	1,211	97.2%	Chi and Caldas (2012)
	Workers and hardhats	N/A	3,320	94.3%	Park et al. (2015)
	Workers and hardhats	N/A	100,000	95.3%	Fang et al. (2018b)
	Workers and hardhats	N/A	3,601	86.5%	Mneymneh et al. (2019)
	Workers, hardhats, harnesses, and anchorages	N/A	33,480	90.0%	Fang et al. (2018a)
Fuzzy inference	Workers and equipment	N/A	1,161	95.6%	Kim et al. (2016)
	Workers and equipment	N/A	300	93.40%	Zhang et al. (2020a)
Fuzzy neural networks	Workers and equipment	1,080	180	91.6%	Hu et al. (2020)
Dual mask translation	Workers and supports	1,461	450	82.5%	Fang et al. (2019)
CNN	Workers, hardhats, and safety vests	1,184	288	67.90%	Nath et al. (2020)
	Workers and harnesses	5,693	77	89.0%	Fang et al. (2018d)
	Workers, hardhats, and eye/face/hand protection	3,652	913	74.4%	Tang et al. (2020b)
DNN	Workers	6,000	50	99.0%	Yan et al. (2019b)
2D–3D geometric projection	Workers and equipment	N/A	1,000	91.0%	Luo et al. (2020a)
	Workers and equipment	N/A	500	0.7 m	Yan et al. (2020)

assessment relies on only the spatial interactions. Hence, to monitor other types of operations, it is difficult to determine whether construction objects are interacting, and what are the natures of their interactions. To address this issue, future studies need to clearly define the concept of interactions and relationships by developing construction ontologies (i.e., a set of *if-then* rules) on a purpose-by-purpose basis. Recently, Zhong et al. (2020) manually built a comprehensive construction ontology for vision-based safety monitoring, and Wang et al. (2019b) exploited crowdsourcing techniques for the development of safety violation ontology. Such ontologies have been applied for safety hazard identification in other studies. Xiong et al. (2019) and Fang et al. (2020d) used a safety-specific ontology when understanding visual relationships and identifying hazardous factors from jobsite images, and Zhang et al. (2020c) evaluated spatial risks between construction workers and equipment by feeding only the object detection results into a safety ontology. Liu et al. (2020) also leveraged an

ontology to understand and describe construction scene images in natural language. These findings indicate that construction ontologies can specify what information is needed for a certain operation (e.g., object types and locations), and can easily transform the visual analysis results to meaningful project information. Thus, ontology-driven approaches can be also developed for productivity monitoring as well as for a wider variety of construction operations.

Another major challenge is to determine how to encode complex interactions into computer-understandable feature vectors. Most research mainly extracted the spatial information (e.g., location, size, velocity) for interaction analysis owing to the encoding difficulty. Fortunately, recent studies shown that the integration of various features can represent the object-to-object interactions more effectively than using only spatial features. Cai et al. (2019) exploited both spatial (e.g., locations, directions) and attentional cues (e.g., worker's head pose) when identifying interacting groups and classifying their activity types. Luo et al. (2020c) also improved the performance of group identification with the additional considerations of deep visual features (extracted by CNN). Further, as indicated by Cai et al. (2019), temporal features can also play a key role in detecting visual relationships of construction objects; this is because the longer observations are available, the better the dynamic interactions among objects can be understood. Moreover, it would be also effective to incorporate operational contexts, which means that a particular type of construction object tends to have a certain type of relationship with a specific type of object. For example, mixer trucks have operational contexts to be "connected to" pump cars rather than to be "next to" when performing a concreting activity. Thus, operational contexts may be helpful for classifying object-to-object relationships more accurately, beyond the spatial interaction (e.g., "next to"). Furthermore, such comprehensive interaction features can be well-learned using GNNs, as reported in recent computer vision research (Hu et al., 2019; Zhou and Chi, 2019).

Multi-Camera-Based Onsite Information Integration and Construction Monitoring

This technology aims to integrate analysis results derived from a set of single cameras, and to obtain all information required for construction monitoring. This objective can be achieved by finding the same object in multiple cameras installed at different physical locations. Park et al. (2012) first presented a stereo-camera-based method that pairs the same worker from 2D images and estimates 3D locations using triangulation theory. The stereo-cameras and triangulation theory were also used to obtain the detailed 3D postures of excavators in one study (Yuan et al., 2017). Soltani et al. (2018) further integrated the stereo systems with location sensors (e.g., global positioning systems) for 3D pose estimation of excavators. However, as it is not common to install stereo-cameras on actual jobsites, Konstantinou and Brilakis (2018) proposed an approach that matches construction workers from four different monocular cameras in indoor environments. The approach calculated the similarity scores for all possible candidates based on their motion, geometry, and color information. In another study (Wei et al., 2019), they used similar features but trained more powerful deep learning models. Zhang et al. (2018a) also developed a vision-based method that simultaneously matches various construction objects, such as workers, excavators, and traffic cones. In a recent study (Kim and Chi, 2020), they proposed a novel approach for pairing the same dump truck in multiple non-overlapping cameras.

Existing studies demonstrated promising results for multi-camera-based object matching and onsite information integration. However, a significant challenge remains to be addressed because only a few attempts have been made in this field; only seven journal articles were retrieved. First, previous research focused on finding the same object in cameras having quite large overlapping areas, even though multiple cameras are generally installed while maximizing the visible coverage areas (i.e., with fewer overlapping areas). Therefore, it is a major challenge to find the same object in multiple non-overlapping cameras. Fortunately, construction researchers can refer to a multi-target multi-camera (MTMC) tracking approach that is widely adopted in the computer vision domain. The MTMC tracking aims to track and re-identify multiple objects (e.g., people, cars) across a multi-camera network. With this paradigm, most state-of-the-art algorithms mainly rely on spatio-temporal cues (e.g., location and speed over time) and layout information (e.g., city map). Such approaches can be also effective in construction environments; once a mixer truck arrives at an entry zone, it may reappear in a concreting zone after a certain period of time (**Figure 7**). In this sense, recent algorithms, developed by in existing computer vision studies (Ristani and Tomasi, 2018; Hsu et al., 2019; Tang et al., 2019), can be examined, and the experimental results can play a crucial role in deriving critical technical issues and providing development guidelines in the area of construction object re-identification. Building on these findings, the incorporation of construction-specific contexts is recommended. In one study (Kim and Chi, 2020), their method effectively paired the same dump truck by considering queueing disciplines of earthmoving operations from non-overlapping cameras: dump trucks are loaded by an excavator in a specific arrival sequence, which means that the first-in dump truck in an entry zone may be the same as the first-served entity in a loading zone. These jobsite- and object-specific characteristics would be beneficial for multi-camera-based object matching and onsite information integration. In the meantime, it may be also reasonable to consider current practices for which construction firms often mark unique IDs, such as QR codes or numbering on hardhats, on onsite workers and equipment. As its feasibility was confirmed by Azar (2015), the detection of QR codes on construction equipment may be a realistic solution.

OPEN RESEARCH CHALLENGES AND FUTURE DIRECTIONS

This section proposes open research challenges and future directions for operation-level vision-based monitoring on construction sites. In particular, theoretical and practical issues

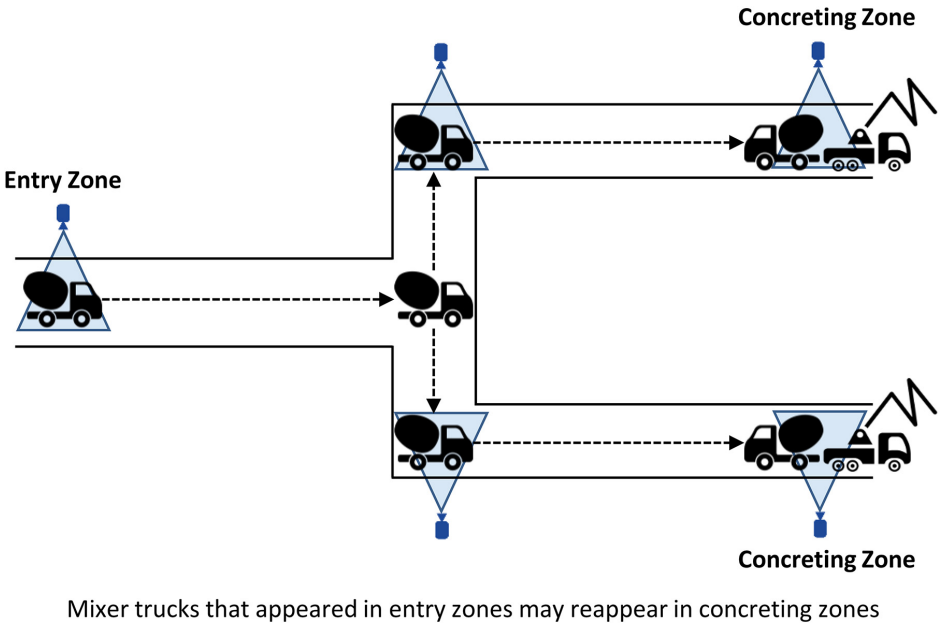

FIGURE 7 | Spatio-temporal cues and site-layout information for multi-target multi-camera tracking.

when applying computer vision technologies to real-world construction environments are discussed.

Benchmark Dataset

As claimed in many existing literatures (Seo et al., 2015a; Zhong et al., 2019; Fang et al., 2020b,c; Zhang et al., 2020b), it is necessary to build and share a comprehensive construction benchmark dataset for model development and validation. While researchers in the computer vision field established diverse benchmark datasets (e.g., ImageNet, Activity Net, MS COCO), construction researchers validated their methods using different datasets. Even if a few studies already developed open construction datasets, each dataset can be used only for a certain technology, such as object detection (Tajeen and Zhu, 2014; Kim et al., 2018b) and action recognition (Luo et al., 2018c, 2020c; Roberts and Golparvar-Fard, 2019), hindering the comprehensive validation of vision-based systems. Therefore, it is difficult to compare model performances fairly and to identify technical problems that commonly arise in many algorithms. The author suggests to develop purpose-specific benchmark datasets for end-to-end framework validation. For example, for the purpose of earthmoving productivity monitoring, benchmark datasets should enable the simultaneous testing of the performance of object detection, action recognition, and interaction analysis. Further, when expanding to multi-camera-based construction monitoring, site-layout and actual productivity values (e.g., input resources, hourly work amount, schedule) should be also embedded together. To this end, it would be effective to form a basis dataset using existing ones, apply crowdsourcing techniques (Liu and Golparvar-Fard, 2015; Han and Golparvar-Fard, 2017; Wang et al., 2019b), and utilize convenient annotation tools (Roberts et al., 2019a,b).

Database-Free Vision-Based Monitoring

As most state-of-the-art technologies originate from traditional deep learning algorithms, it is necessary to build an extensive and high-quality training image database (DB). This requires the manual annotation of construction objects and/or their operational information (e.g., object types and locations) on every single image frame, which is extremely labor-intensive, time-consuming, and expensive. Hence, several efforts have been made to generate synthetic images from virtual models (e.g., 3D CAD, BIM) (Soltani et al., 2016; Braun and Borrmann, 2019) and to oversample given training data using generative adversarial networks (Bang et al., 2020). Many computer vision studies are also underway to minimize the amount of training data and human effort while maximizing the model performance. For example, Roy et al. (2018) and Brust et al. (2019) employed active learning algorithms to train an object detector. As active learning algorithms select the most informative-to-learn instances from abundant unlabeled training data and train a deep learning model with the selected data first, it is possible to significantly reduce the amount of training data and the human effort required for DB development (Kim et al., 2020b). Other researchers also investigated the few-shot learning algorithm (Kang et al., 2019; Wang et al., 2019c; Yan et al., 2019a), which aims to learn and detect new types of target objects even if a small amount of training data are given, i.e., less than 30 images. This would be very beneficial for construction sites where diverse types of construction resources exist and they often vary from phase to phase. Future research can exploit the aforementioned approaches and comparatively analyze their performances. The comprehensive results can build a solid foundation and provide valuable insights for DB-free vision-based monitoring on construction sites.

Visual Data Overload

With the increasing availability of CCTVs, a large amount of visual data (e.g., still images, videos) is being collected, resulting in an inevitable problem: visual data overload. When recording a video-stream with a 1280 × 720 resolution and 10 fps, the amount of video data collected from one camera for eight working hours is about 33.0 GB (Seagate, 2020). Assuming that a project proceeds for 1 year (i.e., 365 days) with five cameras installed, 60 TB of data storage are required, and the size of visual data may even reach the order of PB for a long-term and large-scale project. To address this issue, several studies attempted to filter out less meaningful images (e.g., images having no objects-of-interest) and store only informative ones (e.g., images having workers who are performing a concreting activity). Chen and Wang (2017) developed a construction video summary system that screens out redundant images based on color, gradient, and texture features, and Ham and Kamari (2019) considered worthy scenes as image frames where objects-of-interest (e.g., workers, equipment, materials) exist. Despite their encouraging findings, it remains a major challenge to define appropriate meaningful images for diverse monitoring purposes. For example, for safety monitoring, it may be also valuable if saving and analyzing people-less images. Accordingly, it is recommended to establish important criteria according to each purpose, and to develop appropriate methodologies. With further studies, it will be possible to address data overload issues and facilitate visual analytics.

Integration With Other Technologies

It would be favorable to integrate computer vision with other technologies to address intrinsic shortcomings of ground-level digital cameras. First, location and identity sensors, such as the global positioning system, radio frequency identification (RFID), and Bluetooth low energy (BLE), can be used to counterbalance the main failure cases of vision-based analysis, i.e., occlusions and disappearances. Previous research confirmed the potential usefulness of location and identity sensors. Soltani et al. (2018) proposed a fusion of computer vision and RFID for 3D pose estimation of hydraulic excavators, and Cai and Cai (2020) also combined BLE sensors to obtain the 3D spatial locations of construction workers from onsite images. Second, the integration with construction simulation would be also advantageous. While sensing technologies, including computer vision, can capture the current status of construction sites and resources (e.g., locations, working states), simulation techniques (e.g., discrete event simulation, system dynamics, agent-based modeling) enable the modeling and analysis of the complex relationships involved in construction processes. In this regard, operational information extracted by computer vision (e.g., activity durations, number of input resources) can be used as the inputs of construction simulations. Several studies shown the high potential of such imaging-to-simulation approaches. Kim et al. (2018a, 2019b) fed the results of vision-based object detection to a construction simulation model for the analysis of earthwork productivity, and Golabchi et al. (2018) developed a computer vision-simulation integrated framework for workers' ergonomic risk assessment. Based on these findings, further research can explore the applicability to different construction projects and operations. Finally, it would be possible to mount digital cameras on UAVs to periodically acquire top-view images. As previous research confirmed the usability of camera-equipped UAVs (Kim et al., 2019a, 2020a; Bang and Kim, 2020; Bang et al., 2020), project managers can gather useful information that are not acquired from ground-level images. Regarding the productivity, UAV-acquired images can provide information about the performance of output production (e.g., volume of soil excavated) (Siebert and Teizer, 2014). In other studies (Kim et al., 2019a, 2020a), they monitored the spatial trajectory and proximity of construction objects using UAV images to prevent contact-driven safety accidents.

CONCLUSION

This study comprehensively reviewed existing literatures, derived major research challenges, and proposed future directions for operation-level vision-based monitoring on construction sites. Total 119 papers were thoroughly examined and discussed in the aspect of operation-level monitoring technologies: (1) camera placement for operation-level construction monitoring, (2) single-camera-based construction monitoring and documentation, and (3) multi-camera-based onsite information integration and construction monitoring. Research trends, major open challenges, and future directions were described in detail for each technology. Theoretical and practical issues that may arise when applying computer vision techniques to real-world construction sites were also discussed. Furthermore, cutting-edge algorithms presented in top-tier computer vision conferences were also introduced to indicate potential solutions for research challenges. The findings of this study can form the basis of future research and facilitate the implementation of vision-based construction monitoring and documentation.

AUTHOR CONTRIBUTIONS

JK performed the research and wrote the manuscript.

FUNDING

This research was supported by a grant (20CTAP-C151784-02) from Technology Advancement Research Program funded by Ministry of Land, Infrastructure and Transport of Korean Government.

ACKNOWLEDGMENTS

The author would like to thank the editor and reviewers for their valuable and insightful comments on this manuscript.

REFERENCES

Albahri, A. H., and Hammad, A. (2017a). A novel method for calculating camera coverage in buildings using BIM. *J. Inform. Technol. Constr.* 22, 16–33.

Albahri, A. H., and Hammad, A. (2017b). Simulation-based optimization of surveillance camera types, number, and placement in buildings using BIM. *J. Comp. Civil Eng.* 31:04017055. doi: 10.1061/(ASCE)CP.1943-5487.0000704

Angah, O., and Chen, A. Y. (2020). Tracking multiple construction workers through deep learning and the gradient based method with re-matching based on multi-object tracking accuracy. *Automat. Constr.* 119:103308. doi: 10.1016/j.autcon.2020.103308

Arabi, S., Haghighat, A., and Sharma, A. (2020). A deep-learning-based computer vision solution for construction. *Comp. Aided Civil Infrastruct. Eng.* 35, 753–767. doi: 10.1111/mice.12530

Asadzadeh, A., Arashpour, M., Li, H., Ngo, T., Bab-Hadiashar, A., and Rashidi, A. (2020). Sensor-based safety management. *Automat. Constr.* 113:103128. doi: 10.1016/j.autcon.2020.103128

Azar, E. R. (2015). Construction equipment identification using marker-based recognition and an active zoom camera. *J. Comp. Civil Eng.* 30:04015033. doi: 10.1061/(ASCE)CP.1943-5487.0000507

Azar, E. R., Dickinson, S., and McCabe, B. (2013). Server-customer interaction tracker: computer vision–based system to estimate dirt-loading cycles. *J. Constr. Eng. Manag.* 139, 785–794. doi: 10.1061/(ASCE)CO.1943-7862.0000652

Azar, E. R., and McCabe, B. (2012a). Automated visual recognition of dump trucks in construction videos. *J. Comp. Civil Eng.* 26, 769–781. doi: 10.1061/(ASCE)CP.1943-5487.0000179

Azar, E. R., and McCabe, B. (2012b). Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos. *Automat. Constr.* 24, 194–202. doi: 10.1016/j.autcon.2012.03.003

Bang, S., Baek, F., Park, S., Kim, W., and Kim, H. (2020). Image augmentation to improve construction resource detection using generative adversarial networks, cut-and-paste, and image transformation techniques. *Automat. Constr.* 115:103198. doi: 10.1016/j.autcon.2020.103198

Bang, S., and Kim, H. (2020). Context-based information generation for managing UAV-acquired data using image captioning. *Automat. Constr.* 112:103116. doi: 10.1016/j.autcon.2020.103116

Bergmann, P., Meinhardt, T., and Leal-Taixe, L. (2019). "Tracking without bells and whistles," in *Proceedings of the 2019 IEEE International Conference on Computer Vision*, (Seoul: IEEE), 941–951. doi: 10.1109/ICCV.2019.00103

Bochinski, E., Senst, T., and Sikora, T. (2018). "Extending IOU based multi-object tracking by visual information," in *Proceedings of the 2018 IEEE International Conference on Advanced Video and Signal-Based Surveillance*, (Taipei: IEEE), doi: 10.1109/AVSS.2018.8639144

Boton, C. (2018). Supporting constructability analysis meetings with Immersive Virtual Reality-based collaborative BIM 4D simulation. *Automat. Constr.* 96, 1–15. doi: 10.1016/j.autcon.2018.08.020

Braun, A., and Borrmann, A. (2019). Combining inverse photogrammetry and BIM for automated labeling of construction site images for machine learning. *Automat. Constr.* 106:102879. doi: 10.1016/j.autcon.2019.102879

Brilakis, I., Park, M. W., and Jog, G. (2011). Automated vision tracking of project related entities. *Adv. Eng. Inform.* 25, 713–724. doi: 10.1016/j.aei.2011.01.003

Brust, C. A., Käding, C., and Denzler, J. (2019). "Active learning for deep object detection," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, eds A. Tremeau, G. M. Farinella, and J. Braz (Pargue: Springer), 181–190. doi: 10.5220/0007248601810190

Bügler, M., Borrmann, A., Ogunmakin, G., Vela, P. A., and Teizer, J. (2017). Fusion of photogrammetry and video analysis for productivity assessment of earthwork processes. *Comp. Aided Civil Infrastruct. Eng.* 32, 107–123. doi: 10.1111/mice.12235

Cai, J., and Cai, H. (2020). Robust hybrid approach of vision-based tracking and radio-based identification and localization for 3D tracking of multiple construction workers. *J. Comp. Civil Eng.* 34:04020021. doi: 10.1061/(asce)cp.1943-5487.0000901

Cai, J., Zhang, Y., and Cai, H. (2019). Two-step long short-term memory method for identifying construction activities through positional and attentional cues. *Automat. Constr.* 106:102886. doi: 10.1016/j.autcon.2019.102886

Chen, C., Zhu, Z., and Hammad, A. (2020). Automated excavators activity recognition and productivity analysis from construction site surveillance videos. *Automat. Constr.* 110:103045. doi: 10.1016/j.autcon.2019.103045

Chen, H.-T., Wu, S.-W., and Hsieh, S.-H. (2013). Visualization of CCTV coverage in public building space using BIM technology. *Vis. Eng.* 1:5. doi: 10.1186/2213-7459-1-5

Chen, L., and Wang, Y. (2017). Automatic key frame extraction in continuous videos from construction monitoring by using color, texture, and gradient features. *Automat. Constr.* 81, 355–368. doi: 10.1016/j.autcon.2017.04.004

Chen, Q., García de Soto, B., and Adey, B. T. (2018). Construction automation: research areas, industry concerns and suggestions for advancement. *Automat. Constr.* 94, 22–38. doi: 10.1016/j.autcon.2018.05.028

Chen, Y., Wang, H., Sun, X., Fan, B., and Tang, C. (2020). *Deep Attention Aware Feature Learning for Person Re-Identification*. Available online at: http://arxiv.org/abs/2003.00517 (accessed August 25, 2020).

Chi, S., and Caldas, C. H. (2011). Automated object identification using optical video cameras on construction sites. *Comp. Aided Civil Infrastruct. Eng.* 26, 368–380. doi: 10.1111/j.1467-8667.2010.00690.x

Chi, S., and Caldas, C. H. (2012). Image-based safety assessment: automated spatial safety risk identification of earthmoving and surface mining activities. *J. Constr. Eng. Manag.* 138, 341–351. doi: 10.1061/(ASCE)CO.1943-7862.0000438

Chi, S., Caldas, C. H., and Kim, D. Y. (2009). A methodology for object identification and tracking in construction based on spatial modeling and image matching techniques. *Comp. Aided Civil Infrastruct. Eng.* 24, 199–211. doi: 10.1111/j.1467-8667.2008.00580.x

Choi, M., Ahn, S., and Seo, J. (2020). VR-based investigation of forklift operator situation awareness for preventing collision accidents. *Accident Anal. Prevent.* 136:105404. doi: 10.1016/j.aap.2019.105404

Chu, W., Han, S., Luo, X., and Zhu, Z. (2020). Monocular vision–based framework for biomechanical analysis or ergonomic posture assessment in modular construction. *J. Comp. Civil Eng.* 34:04020018. doi: 10.1061/(ASCE)CP.1943-5487.0000897

Ding, L., Fang, W., Luo, H., Love, P. E. D., Zhong, B., and Ouyang, X. (2018). A deep hybrid learning model to detect unsafe behavior: integrating convolution neural networks and long short-term memory. *Automat. Constr.* 86, 118–124. doi: 10.1016/j.autcon.2017.11.002

Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., and Li, C. (2018a). Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment. *Automat. Constr.* 93, 148–164. doi: 10.1016/j.autcon.2018.05.022

Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M., et al. (2018b). Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Automat. Constr.* 85, 1–9. doi: 10.1016/J.AUTCON.2017.09.018

Fang, Q., Li, H., Luo, X., Ding, L., Rose, T. M., An, W., et al. (2018c). A deep learning-based method for detecting non-certified work on construction sites. *Adv. Eng. Inform.* 35, 56–68. doi: 10.1016/j.aei.2018.01.001

Fang, Q., Li, H., Luo, X., Li, C., and An, W. (2020a). A sematic and prior-knowledge-aided monocular localization method for construction-related entities. *Comp.Aided Civil Infrastruct. Eng.* 1, 1–18. doi: 10.1111/mice.12541

Fang, W., Ding, L., Love, P. E. D., Luo, H., Li, H., Peña-Mora, F., et al. (2020b). Computer vision applications in construction safety assurance. *Automat. Constr.* 110:103013. doi: 10.1016/j.autcon.2019.103013

Fang, W., Ding, L., Luo, H., and Love, P. E. D. (2018d). Falls from heights: a computer vision-based approach for safety harness detection. *Automat. Constr.* 91, 53–61. doi: 10.1016/j.autcon.2018.02.018

Fang, W., Ding, L., Zhong, B., Love, P. E. D., and Luo, H. (2018e). Automated detection of workers and heavy equipment on construction sites: a convolutional neural network approach. *Adv. Eng. Inform.* 37, 139–149. doi: 10.1016/j.aei.2018.05.003

Fang, W., Love, P. E. D., Luo, H., and Ding, L. (2020c). Computer vision for behaviour-based safety in construction: a review and future directions. *Adv. Eng. Inform.* 43:100980. doi: 10.1016/j.aei.2019.100980

Fang, W., Ma, L., Love, P. E. D., Luo, H., Ding, L., and Zhou, A. (2020d). Knowledge graph for identifying hazards on construction sites: integrating

computer vision with ontology. *Automat. Constr.* 119:103310. doi: 10.1016/j.autcon.2020.103310

Fang, W., Xue, J., Zhong, B., Xu, S., Zhao, N., Luo, H., et al. (2019). A deep learning-based approach for mitigating falls from height with computer vision: convolutional neural network. *Adv. Eng. Inform.* 39, 170–177. doi: 10.1016/j.aei.2018.12.005

Golabchi, A., Guo, X., Liu, M., Han, S., Lee, S., and AbouRizk, S. (2018). An integrated ergonomics framework for evaluation and design of construction operations. *Automat. Constr.* 95, 72–85. doi: 10.1016/j.autcon.2018.08.003

Golparvar-Fard, M., Heydarian, A., and Niebles, J. C. (2013). Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Adv. Eng. Inform.* 27, 652–663. doi: 10.1016/J.AEI.2013.09.001

Gong, J., and Caldas, C. H. (2010). Computer vision-based video interpretation model for automated productivity analysis of construction operations. *J. Comp. Civil Eng.* 24, 252–263. doi: 10.1061/(ASCE)CP.1943-5487.0000027

Gong, J., and Caldas, C. H. (2011). An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations. *Automat. Constr.* 20, 1211–1226. doi: 10.1016/j.autcon.2011.05.005

Gong, J., Caldas, C. H., and Gordon, C. (2011). Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. *Adv. Eng. Inform.* 25, 771–782. doi: 10.1016/j.aei.2011.06.002

Guo, Y., and Cheung, N. M. (2018). "Efficient and deep person re-identification using multi-level similarity," in *Proceedings of the 2018 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, UT: IEEE), 2335–2344. doi: 10.1109/CVPR.2018.00248

Guo, Y., Xu, Y., and Li, S. (2020). Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network. *Automat. Constr.* 112:103124. doi: 10.1016/j.autcon.2020.103124

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., and Alahi, A. (2018). "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the 2018 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, UT: IEEE), 2255–2264. doi: 10.1109/CVPR.2018.00240

Ham, Y., Han, K. K., Lin, J. J., and Golparvar-Fard, M. (2016). Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): a review of related works. *Vis. Eng.* 4, 1–8. doi: 10.1186/s40327-015-0029-z

Ham, Y., and Kamari, M. (2019). Automated content-based filtering for enhanced vision-based documentation in construction toward exploiting big visual data from drones. *Automat. Constr.* 105:102831. doi: 10.1016/j.autcon.2019.102831

Han, K., and Golparvar-Fard, M. (2017). Crowdsourcing BIM-guided collection of construction material library from site photologs. *Vis. Eng.* 5:14. doi: 10.1186/s40327-017-0052-3

Han, S., Achar, M., Lee, S., and Peña-Mora, F. (2013a). Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring. *Vis. Eng.* 1:6. doi: 10.1186/2213-7459-1-6

Han, S., and Lee, S. (2013). A vision-based motion capture and recognition framework for behavior-based safety management. *Automat. Constr.* 35, 131–141. doi: 10.1016/J.AUTCON.2013.05.001

Han, S., Lee, S., and Peña-mora, F. (2013b). Vision-based detection of unsafe actions of a construction worker: case study of ladder climbing. *J. Comp. Civil Eng.* 27, 635–644. doi: 10.1061/(ASCE)CP.1943-5487.0000279

Han, S., Lee, S., and Peña-Mora, F. (2014). Comparative study of motion features for similarity-based modeling and classification of unsafe actions in construction. *J. Comp. Civil Eng.* 28:A4014005. doi: 10.1061/(ASCE)CP.1943-5487.0000339

Hsu, H., Huang, T., Wang, G., Cai, J., Lei, Z., and Hwang, J. (2019). "Multi-camera tracking of vehicles based on deep features Re-ID and trajectory-based camera link models," in *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, (Long Beach, CA: IEEE), 416–424. doi: 10.1109/CVPR.2019.00007

Hu, Q., Bai, Y., He, L., Cai, Q., Tang, S., Ma, G., et al. (2020). Intelligent framework for worker-machine safety assessment. *J. Constr. Eng. Manag.* 146:04020045. doi: 10.1061/(ASCE)CO.1943-7862.0001801

Hu, Y., Chen, S., Chen, X., Zhang, Y., and Gu, X. (2019). "Neural message passing for visual relationship detection," in *Proceedings of the 2019 International Conference on Machine Learning*, Long Beach, CA.

Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., and Darrell, T. (2019). "Few-shot object detection via feature reweighting," in *Proceedings of the 2019 IEEE International Conference on Computer Vision 2019*, (Seoul: IEEE).

Kim, D., Lee, S., and Kamat, V. R. (2020a). Proximity prediction of mobile objects to prevent contact-driven accidents in co-robotic construction. *J. Comp. Civil Eng.* 34:04020022. doi: 10.1061/(asce)cp.1943-5487.0000899

Kim, D., Liu, M., Lee, S., and Kamat, V. R. (2019a). Remote proximity monitoring between mobile construction resources using camera-mounted UAVs. *Automat. Constr.* 99, 168–182. doi: 10.1016/j.autcon.2018.12.014

Kim, H., Bang, S., Jeong, H., Ham, Y., and Kim, H. (2018a). Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation. *Automat. Constr.* 92, 188–198. doi: 10.1016/j.autcon.2018.04.002

Kim, H., Ham, Y., Kim, W., Park, S., and Kim, H. (2019b). Vision-based nonintrusive context documentation for earthmoving productivity simulation. *Automat. Constr.* 102, 135–147. doi: 10.1016/j.autcon.2019.02.006

Kim, H., Kim, H., Hong, Y. W., and Byun, H. (2018b). Detecting construction equipment using a region-based fully convolutional network and transfer learning. *J. Comp. Civil Eng.* 32:04017082. doi: 10.1061/(ASCE)CP.1943-5487.0000731

Kim, H., Kim, K., and Kim, H. (2016). Vision-based object-centric safety assessment using fuzzy inference: monitoring struck-by accidents with moving objects. *J. Comp. Civil Eng.* 30:04015075. doi: 10.1061/(ASCE)CP.1943-5487.0000562

Kim, J. (2019). *Automated Construction Site Monitoring: Multi Vision-Based Operational Context Analysis for Enhancing Earthmoving Productivity*. Available online at: http://hdl.handle.net/10371/161874 (accessed May 14, 2020).

Kim, J., and Chi, S. (2017). Adaptive detector and tracker on construction sites using functional integration and online learning. *J. Comp. Civil Eng.* 31:04017026. doi: 10.1061/(ASCE)CP.1943-5487.0000677

Kim, J., and Chi, S. (2019). Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles. *Automat. Constr.* 104, 255–264. doi: 10.1016/j.autcon.2019.03.025

Kim, J., and Chi, S. (2020). Multi-camera vision-based productivity monitoring of earthmoving operations. *Automat. Constr.* 112:103121. doi: 10.1016/j.autcon.2020.103121

Kim, J., Chi, S., and Choi, M. (2019c). "Sequential pattern learning of visual features and operation cycles for vision-based action recognition of earthmoving excavators," in *Computing in Civil Engineering 2019: Data, Sensing, and Analytics*, (Atlanta, GA: American Society of Civil Engineers), 298–304. doi: 10.1061/9780784479247.083

Kim, J., Chi, S., and Hwang, B.-G. (2017). "Vision-based activity analysis framework considering interactive operation of construction equipment," in *ASCE International Workshop on Computing in Civil Engineering* 2017, (Seattle, WA: American Society of Civil Engineers), 162–170. doi: 10.1061/9780784480830.021

Kim, J., Chi, S., and Seo, J. (2018c). Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks. *Automat. Constr.* 87, 297–308. doi: 10.1016/J.AUTCON.2017.12.016

Kim, J., Ham, Y., Chung, Y., and Chi, S. (2018d). "Camera placement optimization for vision-based monitoring on construction sites," in *Proceedings of the 35th International Symposium on Automation and Robotics in Construction*, (Berlin: International Association for Automation and Robotics in Construction), 748–752. doi: 10.22260/ISARC2018/0102

Kim, J., Ham, Y., Chung, Y., and Chi, S. (2019d). Systematic camera placement framework for operation-level visual monitoring on construction jobsites. *J. Constr. Eng. Manag.* 145:04019019. doi: 10.1061/(ASCE)CO.1943-7862.0001636

Kim, J., Hwang, J., Chi, S., and Seo, J. (2020b). Towards database-free vision-based monitoring on construction sites: a deep active learning approach. *Automat. Constr.* 120:103376. doi: 10.1016/j.autcon.2020.103376

Konstantinou, E., and Brilakis, I. (2018). Matching construction workers across views for automated 3D vision tracking on-site. *J. Constr. Eng. Manag.* 144:04018061. doi: 10.1061/(ASCE)CO.1943-7862.0001508

Konstantinou, E., Lasenby, J., and Brilakis, I. (2019). Adaptive computer vision-based 2D tracking of workers in complex environments. *Automat. Constr.* 103, 168–184. doi: 10.1016/j.autcon.2019.01.018

Korea Construction Technology Promotion Act (2016). *Enforcement decree article 98 and 99, statutes of the Republic of Korea*. Available online at: http://law.go.kr/법령/건설기술관리법 (accessed January 28, 2019).

Li, X., Yi, W., Chi, H. L., Wang, X., and Chan, A. P. C. (2018). A critical review of virtual and augmented reality (VR/AR) applications in construction safety. *Automat. Constr.* 86, 150–162. doi: 10.1016/j.autcon.2017.11.003

Liang, C., Lundeen, K. M., Mcgee, W., Menassa, C. C., and Lee, S. (2019). A vision-based marker-less pose estimation system for articulated construction robots. *Automat. Constr.* 104, 80–94. doi: 10.1016/j.autcon.2019.04.004

Liang, X., Shen, G. Q., and Bu, S. (2016). Multiagent systems in construction: A ten-year review. *J. Comp. Civil Eng.* 30, 1–11. doi: 10.1061/(ASCE)CP.1943-5487.0000574

Liu, H., Wang, G., Huang, T., He, P., Skitmore, M., and Luo, X. (2020). Manifesting construction activity scenes via image captioning. *Automat. Constr.* 119:103334. doi: 10.1016/j.autcon.2020.103334

Liu, K., and Golparvar-Fard, M. (2015). Crowdsourcing construction activity analysis from jobsite video streams. *J. Constr. Eng. Manag.* 141, 1–19. doi: 10.1061/(ASCE)CO.1943-7862.0001010

Liu, M., Han, S., and Lee, S. (2016). Tracking-based 3D human skeleton extraction from stereo video camera toward an on-site safety and ergonomic analysis. *Constr. Innovat.* 16, 348–367. doi: 10.1108/CI-10-2015-0054

Luo, H., Liu, J., Fang, W., Love, P. E. D., Yu, Q., and Lu, Z. (2020a). Real-time smart video surveillance to manage safety: a case study of a transport mega-project. *Adv. Eng. Inform.* 45:101100. doi: 10.1016/j.aei.2020.101100

Luo, H., Wang, M., Wong, P. K. Y., and Cheng, J. C. P. (2020b). Full body pose estimation of construction equipment using computer vision and deep learning techniques. *Automat. Constr.* 110:103016. doi: 10.1016/j.autcon.2019.103016

Luo, H., Xiong, C., Fang, W., Love, P. E. D., Zhang, B., and Ouyang, X. (2018a). Convolutional neural networks: computer vision-based workforce activity assessment in construction. *Automat. Constr.* 94, 282–289. doi: 10.1016/j.autcon.2018.06.007

Luo, X., Li, H., Cao, D., Dai, F., Seo, J., and Lee, S. (2018b). Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. *J. Comp. Civil Eng.* 32:04018012. doi: 10.1061/(ASCE)CP.1943-5487.0000756

Luo, X., Li, H., Cao, D., Yu, Y., Yang, X., and Huang, T. (2018c). Towards efficient and objective work sampling: recognizing workers' activities in site surveillance videos with two-stream convolutional networks. *Automat. Constr.* 94, 360–370. doi: 10.1016/j.autcon.2018.07.011

Luo, X., Li, H., Yang, X., Yu, Y., and Cao, D. (2019). Capturing and understanding workers' activities in far-field surveillance videos with deep action recognition and bayesian nonparametric learning. *Comp. Aided Civil Infrastruct. Eng.* 34, 333–351. doi: 10.1111/mice.12419

Luo, X., Li, H., Yu, Y., Zhou, C., and Cao, D. (2020c). Combining deep features and activity context to improve recognition of activities of workers in groups. *Comp. Aided Civil Infrastruct. Eng.* 35, 965-978. doi: 10.1111/mice.12538

Memarzadeh, M., Golparvar-Fard, M., and Niebles, J. C. (2013). Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. *Automat. Constr.* 32, 24–37. doi: 10.1016/j.autcon.2012.12.002

Milan, A., Rezatofighi, S. H., Dick, A., Reid, I., and Schindler, K. (2017). "Online multi-target tracking using recurrent neural networks," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, (San Francisco, CA: AAAI), 4225–4232.

Mneymneh, B. E., Abbas, M., and Khoury, H. (2019). Vision-based framework for intelligent monitoring of hardhat wearing on construction sites. *J. Comp. Civil Eng.* 33:04018066. doi: 10.1061/(ASCE)CP.1943-5487.0000813

Mok, K. Y., Shen, G. Q., and Yang, J. (2015). Stakeholder management studies in mega construction projects: a review and future directions. *Int. J. Project Manag.* 33, 446–457. doi: 10.1016/j.ijproman.2014.08.007

Nath, N. D., Behzadan, A. H., and Paal, S. G. (2020). Deep learning for site safety: Real-time detection of personal protective equipment. *Automat. Constr.* 112:103085. doi: 10.1016/j.autcon.2020.103085

Park, M.-W., and Brilakis, I. (2012). Construction worker detection in video frames for initializing vision trackers. *Automat. Constr.* 28, 15–25. doi: 10.1016/j.autcon.2012.06.001

Park, M.-W., and Brilakis, I. (2016). Continuous localization of construction workers via integration of detection and tracking. *Automat. Constr.* 72, 129–142. doi: 10.1016/j.autcon.2016.08.039

Park, M.-W., Elsafty, N., and Zhu, Z. (2015). Hardhat-wearing detection for enhancing on-site safety of construction workers. *J. Constr. Eng. Manag.* 141:04015024. doi: 10.1061/(ASCE)CO.1943-7862.0000974

Park, M.-W., Koch, C., and Brilakis, I. (2012). Three-dimensional tracking of construction resources using an on-site camera system. *J. Comp. Civil Eng.* 26, 541–549. doi: 10.1061/(ASCE)CP.1943-5487.0000168

Park, M.-W., Makhmalbaf, A., and Brilakis, I. (2011). Comparative study of vision tracking methods for tracking of construction site resources. *Automat. Constr.* 20, 905–915. doi: 10.1016/j.autcon.2011.03.007

Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S. C. (2018). "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11213 LNCS*, New York, NY, 407–423. doi: 10.1007/978-3-030-01240-3_25

Ristani, E., and Tomasi, C. (2018). "Features for multi-target multi-camera tracking and re-identification," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, UT: IEEE), 6036–6046. doi: 10.1109/CVPR.2018.00632

Roberts, D., Calderon, W. T., Tang, S., and Golparvar-fard, M. (2020). Vision-based construction worker activity analysis informed by body posture. *J. Comp. Civil Eng.* 34:04020017. doi: 10.1061/(ASCE)CP.1943-5487.0000898

Roberts, D., and Golparvar-Fard, M. (2019). End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level. *Automat. Constr.* 105:102811. doi: 10.1016/j.autcon.2019.04.006

Roberts, D., Wang, M., Torres Calderon, W., and Golparvar-Fard, M. (2019a). "An annotation tool for benchmarking methods for automated construction worker pose estimation and activity analysis," in *Proceedings of the International Conference on Smart Infrastructure and Construction 2019 (ICSIC): Driving data-informed decision-making*, (Atlanta, GA: ICE), 307–313. doi: 10.1680/icsic.64669.307

Roberts, D., Wang, Y., Sabet, A., and Golparvar-Fard, M. (2019b). "Annotating 2D imagery with 3D kinematically configurable assets of construction equipment for training pose-informed activity analysis and safety monitoring algorithms," in *Proceedings of the ASCE International Conference on Computing in Civil Engineering 2019: Visualization, Information Modeling, and Simulation*, (Atlanta, GA: ASCE), 343–352.

Roy, S., Unmesh, A., and Namboodiri, V. P. (2018). " Deep active learning for object detection," in *Proceedings of the British Machine Vision Conference 2018 (BMVC 2018)*, (Newcastle: British Machine Vision Association), 1–12.

Sadeghian, A., Alahi, A., and Savarese, S. (2017). "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *Proceedings of 2017 IEEE International Conference on Computer Vision*, (Venice: IEEE), 300–311. doi: 10.1109/ICCV.2017.41

Seagate (2020). *Video Storage Calculator*. Available online at: https://www.seagate.com/kr/ko/video-storage-calculator/ (accessed June 11, 2020).

Seo, J., Han, S., Lee, S., and Kim, H. (2015a). Computer vision techniques for construction safety and health monitoring. *Adv. Eng. Inform.* 29, 239–251. doi: 10.1016/J.AEI.2015.02.001

Seo, J., Starbuck, R., Han, S., Lee, S., and Armstrong, T. J. (2015b). Motion data-driven biomechanical analysis during construction tasks on sites. *J. Comp. Civil Eng.* 29:B4014005. doi: 10.1061/(ASCE)CP.1943-5487

Shen, H., Huang, L., Huang, C., and Xu, W. (2018). *Tracklet Association Tracker: An End-to-end Learning-Based Association Approach for Multi-object Tracking*. Available online at: http://arxiv.org/abs/1808.01562 (accessed March 13, 2020).

Sherafat, B., Ahn, C. R., Akhavian, R., Behzadan, A. H., Golparvar-Fard, M., Kim, H., et al. (2020). Automated methods for activity recognition of construction workers and equipment: state-of-the-art review. *J. Constr. Eng. Manag.* 146:03120002. doi: 10.1061/(ASCE)CO.1943-7862.0001843

Siebert, S., and Teizer, J. (2014). Mobile 3D mapping for surveying earthwork projects using an Unmanned Aerial Vehicle (UAV) system. *Automat. Constr.* 41, 1–14. doi: 10.1016/j.autcon.2014.01.004

Soltani, M. M., Zhu, Z., and Hammad, A. (2016). Automated annotation for visual recognition of construction resources using synthetic images. *Automat. Constr.* 62, 14–23. doi: 10.1016/j.autcon.2015.10.002

Soltani, M. M., Zhu, Z., and Hammad, A. (2017). Skeleton estimation of excavator by detecting its parts. *Automat. Constr.* 82, 1–15. doi: 10.1016/j.autcon.2017.06.023

Soltani, M. M., Zhu, Z., and Hammad, A. (2018). Framework for location data fusion and pose estimation of Excavators using stereo vision. *J. Comp. Civil Eng.* 32:04018045. doi: 10.1061/(ASCE)CP.1943-5487.0000783

Son, H., Choi, H., Seong, H., and Kim, C. (2019). Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks. *Automat. Constr.* 99, 27–38. doi: 10.1016/j.autcon.2018.11.033

Swallow, M., and Zulu, S. (2019). Benefits and barriers to the adoption of 4d modeling for site health and safety management. *Front. Built Environ.* 4:86. doi: 10.3389/fbuil.2018.00086

Tajeen, H., and Zhu, Z. (2014). Image dataset development for measuring construction equipment recognition performance. *Automat. Constr.* 48, 1–10. doi: 10.1016/j.autcon.2014.07.006

Tang, S., Golparvar-fard, M., Naphade, M., and Gopalakrishna, M. M. (2020a). Video-based motion trajectory forecasting method for proactive construction safety monitoring systems. *J. Comp. Civil Eng.* 34:04020041. doi: 10.1061/(ASCE)CP.1943-5487.0000923

Tang, S., Roberts, D., and Golparvar-Fard, M. (2020b). Human-object interaction recognition for automatic construction site safety inspection. *Automat. Constr.* 120:103356. doi: 10.1016/j.autcon.2020.103356

Tang, Z., Naphade, M., Xiaodong, M. L., Stan, Y., Wang, S., Kumar, R., et al. (2019). "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, (Long Beach: IEEE), 8797–8806.

Teizer, J. (2015). Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. *Adv. Eng. Inform.* 29, 225–238. doi: 10.1016/J.AEI.2015.03.006

Teizer, J., and Vela, P. A. (2009). Personnel tracking on construction sites using video cameras. *Adv. Eng. Inform.* 23, 452–462. doi: 10.1016/j.aei.2009.06.011

Wang, Q., Guo, Z., Mintah, K., Li, Q., Mei, T., and Li, P. (2019a). Cell-based transport path obstruction detection approach for 4D BIM construction planning. *J. Constr. Eng. Manag.* 145:04018141. doi: 10.1061/(ASCE)CO.1943-7862.0001583

Wang, Y., Liao, P. C., Zhang, C., Ren, Y., Sun, X., and Tang, P. (2019b). Crowdsourced reliable labeling of safety-rule violations on images of complex construction scenes for advanced vision-based workplace safety. *Adv. Eng. Inform.* 42:101001. doi: 10.1016/j.aei.2019.101001

Wang, Y. X., Ramanan, D., and Hebert, M. (2019c). "Meta-learning to detect rare objects," in *Proceedings of 2019 IEEE International Conference on Computer Vision*, (Seoul: IEEE), 9924–9933. doi: 10.1109/ICCV.2019.01002

Wei, R., Love, P. E. D., Fang, W., Luo, H., and Xu, S. (2019). Recognizing people's identity in construction sites with computer vision: a spatial and temporal attention pooling network. *Adv. Eng. Inform.* 42:100981. doi: 10.1016/j.aei.2019.100981

Xiang, Y., Alahi, A., and Savarese, S. (2015). "Learning to track: Online multi-object tracking by decision making," in *Proceedings of 2015 IEEE International Conference on Computer Vision*, (Las Condes: IEEE), 4705–4713. doi: 10.1109/ICCV.2015.534

Xiao, B., and Zhu, Z. (2018). Two-dimensional visual tracking in construction scenarios: a comparative study. *J. Comp. Civil Eng.* 32:04018006. doi: 10.1061/(asce)cp.1943-5487.0000738

Xiong, R., Song, Y., Li, H., and Wang, Y. (2019). Onsite video mining for construction hazards identification with visual relationships. *Adv. Eng. Inform.* 42, 100966. doi: 10.1016/j.aei.2019.100966

Xu, Y., and Brownjohn, J. M. W. (2018). Review of machine-vision based methodologies for displacement measurement in civil structures. *J. Civil Struct. Health Monit.* 8, 91–110. doi: 10.1007/s13349-017-0261-4

Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., and Lin, L. (2019a). "Meta R-CNN: Towards general solver for instance-level low-shot learning," in *Proceedings of the 2019 IEEE International Conference on Computer Vision*, (Seoul: IEEE), 9576–9585. doi: 10.1109/ICCV.2019.00967

Yan, X., Li, H., Wang, C., Seo, J., Zhang, H., and Wang, H. (2017). Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion. *Adv. Eng. Inform.* 34, 152–163. doi: 10.1016/j.aei.2017.11.001

Yan, X., Zhang, H., and Li, H. (2019b). Estimating worker-centric 3D spatial crowdedness for construction safety management using a single 2D camera. *J. Comp. Civil Eng.* 33:04019030. doi: 10.1061/(ASCE)CP.1943-5487.0000844

Yan, X., Zhang, H., and Li, H. (2020). Computer vision-based recognition of 3D spatial relationship between moving objects for monitoring struck-by accidents. *Comp. Aided Civil Infrastruct. Eng.* 35, 1023–1036. doi: 10.1111/mice.12536

Yang, J., Park, M. W., Vela, P. A., and Golparvar-Fard, M. (2015). Construction performance monitoring via still images, time-lapse photos, and video streams: now, tomorrow, and the future. *Adv. Eng. Inform.* 29, 211–224. doi: 10.1016/j.aei.2015.01.011

Yang, J., Shi, Z., and Wu, Z. (2016). Vision-based action recognition of construction workers using dense trajectories. *Adv. Eng. Inform.* 30, 327–336. doi: 10.1016/j.aei.2016.04.009

Yang, J., Vela, P., Teizer, J., and Shi, Z. (2014). Vision-based tower crane tracking for understanding construction activity. *J. Comp. Civil Eng.* 28, 103–112. doi: 10.1061/(ASCE)CP.1943-5487.0000242

Yang, X., Wang, F., and Wang, C. (2018). Computer-aided optimization of surveillance cameras placement on construction sites. *Comp. Aided Civil Infrastruct. Eng.* 33, 1110–1126. doi: 10.1111/mice.12385

Yi, W., and Chan, A. P. C. (2014). Critical review of labor productivity research in construction journals. *J. Manag. Eng.* 30, 214–225. doi: 10.1061/(ASCE)ME.1943-5479.0000194

Yu, Y., Li, H., Umer, W., Dong, C., Yang, X., Skitmore, M., et al. (2019a). Automatic biomechanical workload estimation for construction workers by computer vision and smart insoles. *J. Comp. Civil Eng.* 33:04019010. doi: 10.1061/(ASCE)CP.1943-5487.0000827

Yu, Y., Li, H., Yang, X., Kong, L., Luo, X., and Wong, A. Y. L. (2019b). An automatic and non-invasive physical fatigue assessment method for construction workers. *Automat. Constr.* 103, 1–12. doi: 10.1016/j.autcon.2019.02.020

Yuan, C., Li, S., and Cai, H. (2017). Vision-based excavator detection and tracking using hybrid kinematic shapes and key nodes. *J. Comp. Civil Eng.* 31:04016038. doi: 10.1061/(ASCE)CP.1943-5487.0000602

Zhang, B., Zhu, Z., Hammad, A., and Aly, W. (2018a). Automatic matching of construction onsite resources under camera views. *Automat. Constr.* 91, 206–215. doi: 10.1016/J.AUTCON.2018.03.011

Zhang, H., Yan, X., and Li, H. (2018b). Ergonomic posture recognition using 3D view-invariant features from single ordinary camera. *Automat. Constr.* 94, 1–10. doi: 10.1016/j.autcon.2018.05.033

Zhang, M., Cao, Z., Yang, Z., and Zhao, X. (2020a). Utilizing computer vision and fuzzy inference to evaluate level of collision safety for workers and equipment in a dynamic environment. *J. Construct. Eng. Manag.* 146:04020051. doi: 10.1061/(ASCE)CO.1943-7862.0001802

Zhang, M., Shi, R., and Yang, Z. (2020b). A critical review of vision-based occupational health and safety monitoring of construction site workers. *Safety Sci.* 126:104658. doi: 10.1016/j.ssci.2020.104658

Zhang, M., Zhu, M., and Zhao, X. (2020c). Recognition of high-risk scenarios in building construction based on image semantics. *J. Comp. Civil Eng.* 34:04020019. doi: 10.1061/(ASCE)CP.1943-5487.0000900

Zhang, Y., Luo, H., Skitmore, M., Li, Q., and Zhong, B. (2019). Optimal camera placement for monitoring safety in metro station construction work. *J. Constr. Eng. Manag.* 145, 1–13. doi: 10.1061/(ASCE)CO.1943-7862.0001584

Zhong, B., Li, H., Luo, H., Zhou, J., Fang, W., and Xing, X. (2020). Ontology-based semantic modeling of knowledge in construction: classification and identification of hazards implied in images. *J. Constr. Eng. Manag.* 146, 1–15. doi: 10.1061/(ASCE)CO.1943-7862.0001767

Zhong, B., Wu, H., Ding, L., Love, P. E. D., Li, H., Luo, H., et al. (2019). Mapping computer vision research in construction: developments, knowledge gaps and implications for research. *Automat. Constr.* 107:102919. doi: 10.1016/j.autcon.2019.102919

Zhou, P., and Chi, M. (2019). "Relation parsing neural network for human-object interaction detection," in *Proceedings of 2019 IEEE International Conference on Computer Vision*, (Seoul: IEEE), 843–851. doi: 10.1109/ICCV.2019.00093

Zhu, Z., Ren, X., and Chen, Z. (2016). Visual tracking of construction jobsite workforce and equipment with particle filtering. *J. Comp. Civil Eng.* 30, 1–15. doi: 10.1061/(ASCE)CP.1943-5487.0000573

Zhu, Z., Ren, X., and Chen, Z. (2017). Integrated detection and tracking of workforce and equipment from construction jobsite videos. *Automat. Constr.* 81, 161–171. doi: 10.1016/j.autcon.2017.05.005

Zou, J., and Kim, H. (2007). Using hue, saturation, and value color space for hydraulic excavator idle time analysis. *J. Comp. Civil Eng.* 21, 238–246. doi: 10.1061/(ASCE)0887-3801200721:4(238)

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with JK.