www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

COMPUTER VISION APPLICATIONS IN CONSTRUCTION AND ASSET MANAGEMENT PHASES: A LITERATURE REVIEW

SUBMITTED: May 2021
REVISED: November 2022
PUBLISHED: April 2023
EDITOR: Robert Amor
DOI: 10.36680/j.itcon.2023.009

Zhouqian Jiang, Ph.D. Candidate
Department of Architectural Engineering, The Pennsylvania State University, University Park, PA, USA; zpj5026@psu.edu

John I. Messner, Ph.D., Charles and Elinor Matts Professor
Department of Architectural Engineering, The Pennsylvania State University, University Park, PA, USA; jim101@psu.edu

SUMMARY*: Recent advances in digital photography and unmanned aerial vehicle (UAV) platforms make visual data from construction project sites more accessible to project teams. To semi-automatically or automatically obtain the essential information, evaluate the ongoing activities or operations, and address project-level challenges, researchers have focused on applying various computer vision (CV)-based methods to process and interpret the acquired visual data. This research developed a framework to summarize the vision-based methods that have been applied to construction/asset management operations through a systematic literature review. The reviewed literature was composed of 103 journal papers from 2011 to 2020. All the reviewed journal papers were from the Ei Compendex database with specific search criteria. The developed framework consisted of two parts: use cases and CV domains. Use cases contained five aspects: safety monitoring, productivity improvement, progress monitoring, infrastructure inspection, and robotic application. CV domains contained six aspects: image processing, object classification, object detection, object tracking, pose estimation, and 3D reconstruction. All eleven aspects were integrated from the reviewed papers. For each reviewed paper, the general workflow of applied vision-based approaches was described and categorized into each use case. A trending timeline was developed to analyze the popularity of the identified use cases and CV domains within the reviewed time period. Both the quantity and variety of construction use cases and CV domains have increased. Challenges and limitations of applying CV-based methods in the construction industry were also identified. This paper provides readers with a summary of how CV-based methods have been used in the construction industry and serve as a reference for future research and development.*

KEYWORDS*: Construction Use Cases, Computer Vision, Literature Review, Trending Timeline*

REFERENCE: *Zhouqian Jiang, John I. Messner (2023). Computer Vision Applications In Construction And Asset Management Phases: A Literature Review. Journal of Information Technology in Construction (ITcon), Vol. 28, pg. 176-199, DOI: 10.36680/j.itcon.2023.009*

1. INTRODUCTION

The construction market in the United States is one of the largest globally. In 2020, the value-added of the construction industry contributed about 4.3% of the gross domestic product (GDP) in the United States (Best 2021). Despite its significant economic importance, the construction industry is one of the least automated and digitized industries worldwide (Merrill 2021). Other industries have leveraged computer vision (CV) approaches to support the automation of tasks throughout their industries. There has been substantial growth in using digital cameras to acquire daily images or videos from construction sites in recent years. However, low usage and timely manual interpretation of the acquired visual data are still common in construction projects. Some researchers are exploring the use of CV to interpret visual data to solve these issues, retrieving necessary information for supporting decision-making on construction projects (Jiang and Messner 2020). The goal of applying CV-based methods in the construction and asset management phases is to automate the tasks that require manual observation and inspection and overcome the issues associated with safety hazards and low productivity. Yet, there is a need to develop a framework of the CV approaches that can be applied to construction use cases to inform future research and development activities.

In this paper, the author conducted a comprehensive and systematic literature review of recent CV applications in the construction and asset management phases. In total, 103 journal papers from 2011 to 2020 were gathered and reviewed from leading journals within the construction field. Based on the existing taxonomy, a framework containing 1) use cases and 2) CV domains was developed. Use cases identified various fields within the construction and asset management phases where researchers have applied CV-based methods. CV domains categorized the sub-domains of CV for each vision-based method that reported studies used. Each reviewed paper was categorized into both use cases and CV domains. A trending timeline was developed to highlight the frequency of construction use cases and applied CV domains within the reviewed time frame. The total number of reviewed papers has been increasing over the years, as well as the variety of covered construction use cases and implemented CV domains. The most and least popular construction use cases and CV domains were also discussed. The reviewed papers identified the challenges and limitations of CV-based methods implementation. Challenges focused on the obstacle of technology implementation and social ethics, while limitations focused on technical limitations.

2. UNDERSTANDING OF COMPUTER VISION

CV is a subfield of Artificial Intelligence. The goal of CV is to understand the content of visual data (images and videos). Typically, this involves developing methods to imitate human vision (Huang 1996). Conventional CV techniques relied on extensive manual effort to design rules-based detectors and handcrafted feature descriptors to classify and detect certain objects from images. Some CV algorithms, such as edge detection, corner detection, and blob detection, use feature detectors to extract corresponding features. The extracted features can then be characterized by feature descriptors that capture the local distribution of specific properties, such as gradient directions. The Scale Invariant Feature Transform (SIFT) (Lowe 2004) is a typical feature descriptor to detect local features from images, and is known to be robust to object rotation and scale variations (Nanni et al. 2017). Another commonly used feature descriptor for object classification on construction sites is the Histogram of Oriented Gradients (HOG) (Dalal and Triggs 2005). This technique counts occurrences of gradient orientation in localized portions of an image as features to classify objects. The handcrafted feature descriptors are commonly paired with a classifier, such as the Support Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2000). The way to perform classification or detection tasks is by using a fixed-size rectangle region that slides across an image. The feature descriptor and classifier will be applied for each of these regions to determine if the region contains the object of interest. This process is known as the sliding window method. However, conventional CV techniques are inflexible. Designing the feature descriptors requires engineering and expert experience, since it is necessary to determine which features are important in each given image (Mahony et al. 2020). It can be cumbersome and inefficient as the number of classes that need to be classified increases.

Advances in machine learning have improved the machine's capability of understanding the input data and, as a result, have enhanced developments in the CV area. Deep learning, a subset of machine learning, has made CV algorithms much more accurate and cost-effective. Deep learning enables the creation of complex networks, known as Convolutional Neural Networks (CNNs). Deep layers in these complex networks act as a set of feature extractors that are independent of any specific classification task (Nanni et al. 2017). This means that deep learning

can automatically extract a set of features learned directly from input images (Bora et al. 2016). The generated networks can then be applied to other unseen images to produce an accurate classification.

The general idea behind deep learning is called "end-to-end" learning, where the deep learning model is given a set of images with annotated classes on them. The network hierarchically extracts multiple levels of features representing the annotated class during the learning process. Low-level features usually contain edges and curves in the images. In contrast, higher-level features can represent the semantics of the images, which can provide greater robustness to intra-class variability (Chan et al. 2015). Even though feature extraction is the primary goal for both conventional CV methods and deep learning, the difference is that features are manually designed by experts beforehand based on class characteristics in the conventional CV methods, in contrast to features that will be learned and extracted automatically by CNNs using input data (Nanni et al. 2017). And in general, deep learning performs much better than traditional algorithms.

Deep learning does not make conventional CV methods obsolete. In some cases, conventional CV methods and deep learning are mixed for better performance. For example, the sliding window and the SIFT algorithm are first used to identify the regions with the object of interest. Then deep learning models are applied to the identified regions for reduced processing time. The Principle Component Analysis (PCA) is used before applying deep learning models to reduce feature size and prevent model overfitting. CV-based methods have been adopted into various domains, including image processing, object classification, object detection, pose estimation, and 3D reconstruction. A brief introduction of each domain will be presented below, as they form the foundation for understanding the application of CV-based methods in the construction and asset management phases.

2.1 Image Processing

Image processing uses a digital computer to process either images or video frames through an algorithm (Gonzalez and Woods, 2007). Although CV overlaps with image processing on basic techniques, and researchers were using these two terms interchangeably (Babatunde et al. 2015), CV is distinct from image processing in the basic workflow. In image processing, images are taken as inputs and processed with certain transformations; the outputs are still images. In CV, the output would be the quantitative or qualitative information extracted from the input visual data. However, image processing methods are usually implemented before the actual CV methods since the enhanced image quality will improve the performance of the information extraction process (Wiley and Lucas 2018). The goal of image processing is to either enhance the visual data for future interpretation or extract certain information from visual data. Typical image processing tasks include enhancement, restoration, registration, and feature extraction.

2.2 Object Classification

Object classification is a fundamental task in CV. It is a process of predicting the probability of the presence of a specific object class in the input images. A well-known classification network is the AlexNet (Krizhevsky et al. 2012). The AlexNet achieved a top-5 error rate (rate of not finding the true label of a given image among its top 5 predictions) of 15.3%, which outperformed the next best result with 26.2%. The performance of the classification network has continued to improve over time by modifying network structure and parameters. The Visual Geometry Group (VGG-16) proposed by Simonyan and Zisserman (2015) was much deeper and had more parameters than AlexNet. There were multiple differences between the AlexNet and VGG-16, including used combinations of inception modules, which contained pooling and convolutions layers at different scales and concatenation operations. It also used 1x1 feature convolutions to reduce the parameter significantly. The VGG-16 also brought a standard that all kernel filters have a size of 3x3, and max poolings should be placed after every 2 or 3 layers of convolutions. The VGG-16 network achieved a 7.3% top-5 error rate. Another breakthrough was the development of GoogleNet by Szegedy et al. (2014). The GoogleNet was able to achieve a 6.67% top-5 error rate.

2.3 Object Detection

Object detection in the CV domain focuses on identifying the semantic features of an object and determining the location of the object in images. The object detection algorithms can be generally categorized into two-stage and one-stage algorithms. The Regional-based Convolutional Neural Networks (R-CNNs) are a family of two-stage algorithms for addressing object detection and localization tasks. The two-stage algorithms usually involve generating object region proposals with deep neural networks, followed by object detection based on features extracted from the proposed regions. The two-stage algorithms achieve better detection results than the one-stage

algorithms, but are also slower. The first R-CNN was proposed by Girshick et al. (2014), combined with the selective search to extract possible objects. While the proposed network achieved great results of 31.4% mean Average Precision (mAP), the training process was arduous, where proposals needed to be generated for each training dataset, and the CNN feature extraction needed to be applied to each one. The Fast R-CNN (Girshick 2015) was developed to overcome this issue. Instead of using CNN to extract features independently, it applied the Region of Interest (RoI) Pooling Layer on the feature map to extract features specific to a given input candidate region. The proposed model was significantly faster in training and detecting objects, yet it still required a set of candidate regions to be generated along with each input image. Ren et al. (2016) proposed the Faster R-CNN network, which was considered state-of-the-art for its accuracy. Faster R-CNN leveraged the Region Proposal Network (RPN) to replace the Selective Search algorithm, significantly reducing the time of generating region proposals and object detection.

On the other hand, one-stage algorithms are much faster and can be applied in real time. The popular family of one-stage algorithms is the You Only Look Once (YOLO) developed by Redmon et al. (2016). This approach leveraged a single neural network trained end to end, where it predicted the bounding boxes and class labels directly on each input image. The YOLO network achieved a much faster detection speed at the expense of accuracy. The YOLOv2 (Redmon and Farhadi, 2016) and YOLOv3 (Redmon and Farhadi, 2018) were subsequently introduced with improvements in the detection performance. Another notable method for object detection is the Single Shot Detector (SSD) proposed by Liu et al. (2016). This approach was designed to achieve real-time object detection without sacrificing detection accuracy. SSD improved process speed by eliminating the need for the RPN. To recover the loss of accuracy, SSD applied multi-scale feature maps and smaller default boxes. It achieved 74.3% mAP on 300 x 300 image size and 76.8% mAP on 512 x 512 image size, respectively.

2.4 Pose Estimation

Human pose estimation aims to determine the spatial location of human joints from images or video frames, or reconstruct the skeleton data using motion capturing hardware (Kitsikidis et al. 2014). It can be challenging since the occlusions, viewpoints, and illumination changes can cause noise.

Li et al. (2019) proposed multi-stage methods to refine the pose estimation process. The multi-stage methods can be either a bottom-up or a top-down pipeline. The bottom-up pipeline first detects individual joints in the images, and then combines them into human poses. Cao et al. (2017) leveraged a set of 2D vector fields to detect the location and orientation of human limbs in images. The detected key points then went through a multi-stage network to associate them together. The top-down pipeline locates people first using human detectors (Li et al. 2018), then performs the single-person pose estimator to predict key points. Wei et al. (2016) leveraged deep neural networks to detect human pose. This work proposed a sequential architecture composed of convolutional networks to implicitly model long-range dependencies between joints.

One of the most popular bottom-up approaches and state-of-the-art for multi-person human pose estimation is called OpenPose (Cao et al. 2019). This approach features real-time and multi-person pose estimation. The benefits of OpenPose are that it is open-sourced, and supports multiple hardware architectures. It can also be embedded into different data acquisition sources, such as cameras, webcams, and CCTV systems. The state-of-the-art of top-down approach is AlphaPose (Fang et al. 2018). The AlphaPose can detect both single and multi-person poses from images or videos. It followed a two-step framework: identifying the human body with bounding boxes, and then pose estimation. The AlphaPose approach was optimized to detect poses even if the bounding boxes were inaccurate.

2.5 3D Reconstruction

3D reconstruction often refers to recovering the 3D structure of the scene from multi-view images. The CV techniques behind this process are Structure from Motion (SfM) (Ullman 1979) and multi-view stereo (MVS) (Goesele et al. 2007). The SfM algorithm was used to estimate image location, orientation, and camera parameters from unordered image collections, which were essential to generate 3D sparse point clouds of a scene. The approach requires images taken from different positions and angles with a certain amount of overlap. However, without further processing, SfM results can be insufficiently detailed and too noisy for high-quality surface reconstruction (Rosnell and Honkavaara 2012)). The MVS technique can filter out noisy data and increase the number of reconstructed points by two or three orders of magnitude (James and Robson 2012), generating dense 3D point clouds or 3D mesh models.

3D reconstruction using deep learning models has attracted increasing attention and has demonstrated impressive results. Han et al. (2021) developed a survey and classified image-based 3D reconstruction methods using deep learning models into volumetric-based, surface-based, and point-based approaches. Although 3D reconstruction was generally used for generic objects, Han discussed three specific 3D reconstruction applications that can leverage prior knowledge to significantly improve the quality of reconstructed objects: 3D human body reconstruction, 3D face reconstruction, and 3D scene parsing. Huang et al. (2018) proposed a volumetric approach for 3D human body reconstruction, which took multiple RGB images and their corresponding camera parameters as inputs, and generated 3D dense fields. The probability of being inside or outside of the human body was calculated for each voxel. This approach was able to reconstruct detailed geometric shapes of human bodies, but it was limited to simple backgrounds. As for 3D face reconstruction, Feng et al. (2018) proposed a method that leveraged the CNN framework to first generate 3D facial curves from various plane images. These curves were then transformed into 3D point clouds, and facial surfaces were generated to fit the points. This proposed 3D face reconstruction approach was robust to different poses, expressions, and illumination conditions. 3D scene parsing first leverages image object detection and recognition to interpret 3D scenes, and then performs 3D reconstruction. Izadinia et al. (2017) proposed a method to recognize objects from indoor scenes using Faster R-CNN, infer room geometry, and interpret object poses and sizes in the room. This method can better handle the occlusion and clutter issue with limited input data.

3. RESEARCH METHOD

This section introduced the outline of the methods used to conduct the systematic literature review, exploring the recent and relevant research focusing on leveraging CV-based methods in construction and asset management phases. As shown in Fig. 1, the literature review was performed with the Ei Compendex database, a comprehensive source of science and engineering literature. The review focused on journal papers in leading journals within the construction field: Automation in Construction, ASCE Journal of Computing in Civil Engineering, Advanced Engineering Informatics, ASCE Journal of Construction Engineering and Management, Computer-Aided Civil and Infrastructure Engineering, and Journal of Information Technology in Construction.

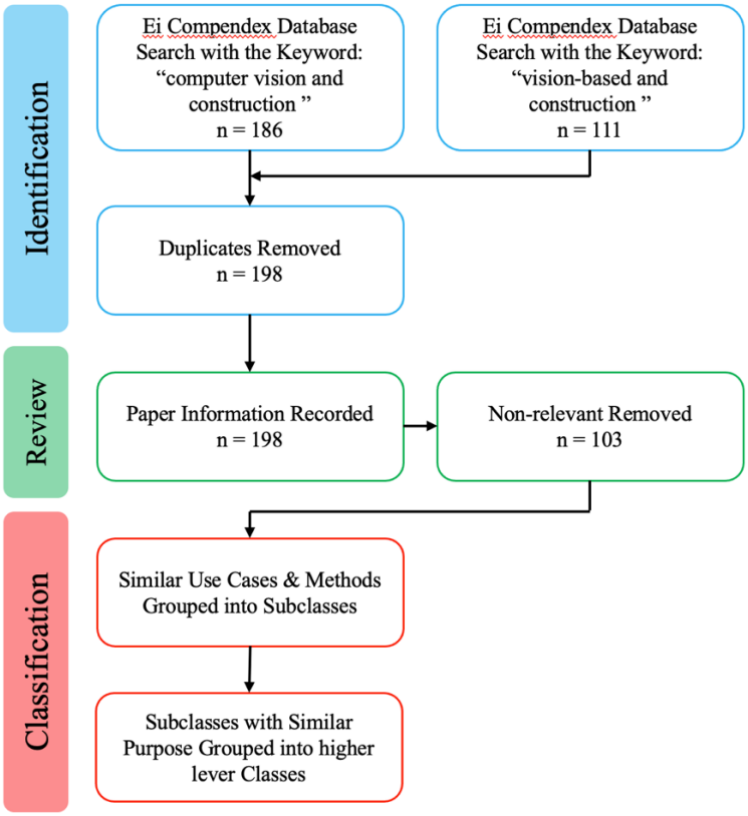

FIG. 1: The process of the systematic literature review.

To identify the studies that have covered both CV and construction fields, two versions of keywords were used for two separate searches: "computer vision and construction" and "vision-based and construction". Journal papers from 2011 to 2020 were considered in this literature review to present how CV methods applied in the construction and asset management phases have evolved over the years while also illustrating the state-of-the-art methods in each CV domain. With the defined search limitations, 186 and 111 journal papers were initially identified from the two separate searches, respectively. The two search streams were merged, and the duplicated papers were removed, resulting in a total of 198 journal papers. A systematic review was then conducted for each identified paper to record information, including use cases, applied CV method(s), and contributions of the reviewed paper. This process could ensure that the relevant studies have been selected. In total, 103 journal papers were chosen for this literature review since the remaining papers did not match the scope of this literature review.

The foundation for categorizing six CV domains was based on the taxonomy of CV domains developed by Khanday and Sofi (2021), including image classification, object detection, object recognition, visual saliency detection, semantic and instance segmentation, human pose estimation, and image retrieval. This existing taxonomy was then modified based on the CV-based methods implemented in the 103 journal papers mentioned above. New domains were added, and similar domains were integrated. The revised CV domains included object classification, object detection, object tracking, pose estimation, 3D reconstruction, and image processing.

The integration process of construction use cases followed the high-level-purpose method identified within the taxonomy by Jiang et al. (2017). All similar use cases were grouped into subclasses. This was fulfilled by looking through all the options and identifying the high-level purpose for which each option would be implemented. These subclasses were then classified into higher classes by repeating the previous grouping method. After the classifying process, five construction use cases were developed. These identified use cases were at the highest classification level and could not be grouped anymore. The categorized CV domains and construction use cases were not comprehensive. They were the categorical systems developed only based on the identified journal papers, which helps readers better understand state-of-the-art CV applications in various construction use cases.

4. RESULTS

The following sections discuss the identified construction use cases where the CV domains were leveraged.

4.1 Safety Monitoring

Safety must always be a priority throughout the construction process. Many researchers have focused on leveraging CV-based methods to ensure safety compliance on sites and prevent work-related injuries among construction workers. Applications of CV-based methods in safety monitoring can be summarized into five aspects: struck-by hazards monitoring, fall hazards prevention, personal protective equipment (PPE) monitoring, safety training, and biomechanical analysis.

Studies have focused on detecting and locating construction entities (Kim et al. 2017, Kim and Chi 2017) and predicting their travel trajectories (Brilakis et al. 2011, Yuan et al. 2017, Lee and Park 2019) to prevent potential struck-by hazards on site. The CV domain that was commonly used to identify and locate construction equipment and workers is object detection. A typical workflow was to apply feature-based descriptors to extract features from visual data. The extracted features were then fed into a classifier to recognize a specific construction entity. The commonly used feature-based descriptor is the Histogram of Oriented Gradients (HOG), which provides better detection results in identifying construction workers and equipment than other descriptors (Rezazadeh Azar and McCabe 2012, Memarzadeh et al. 2013). Support Vector Machine (SVM) is commonly used as the classifier along with the HOG feature descriptor (Park and Brilakis 2016). Rezazadeh Azar and McCabe (2012b) used HOG to train classifiers to detect hydraulic excavators in different poses. Similarly, Kim et al. (2017) leveraged the HOG feature and the SVM classifier to identify construction workers.

Recent studies have focused on applying deep learning models, especially R-CNN, to detect and locate construction entities from visual data. Kim et al. (2018) used a Region-based Fully Convolutional Network (R-FCN) to detect various types of construction equipment. Fang et al. (2018) applied Faster R-CNN to detect construction workers and excavators in real time. The accuracy of the proposed method was 91% for detecting workers and 95% for detecting excavators, which outperformed the conventional handcrafted feature descriptors.

Object tracking methods were also used to analyze the movements of construction entities from video frames while generating trajectories for proactive struck-by hazards prevention. Commonly used tracking methods were the filtering tracking methods, including the Kalman filter-based method (Kalman 1960) and the Particle filter-based method (Gordon et al. 1993). The filtering tracking methods are a top-down process that addresses the dynamics of the tracked objects, awareness of scene priors, and analysis of different hypotheses (Gong and Caldas 2011). Therefore, the filtering tracking methods perform better on complex scenes and objects and can handle occlusion much better. Kim et al. (2016) presented a safety module to monitor the struck-by accidents by tracking moving entities based on Kalman filter-based method and fuzzy inference, while Zhu et al. (2016) leveraged the Particle filter-based method to track workers and moving equipment. Another tracking method was the Kernel-based method (Comaniciu et al. 2003), which has lower computational complexity. Park et al. (2012) and Chen et al. (2017) used the Kernel-based 2D tracking method to identify the 2D pixel coordination of a moving entity's centroid.

Researchers have leveraged object detection methods to identify the presence of edge supporting structures to prevent fall hazards on sites. Fang et al. (2019) used Mask R-CNN (He et al. 2018) to detect and segment supporting structures and people from image backgrounds, then determine the relevant position and relationship between people and supporting structures to identify potential fall hazards. Kolar et al. (2018) applied a CNN-based technique to detect safety guardrails on sites, achieving approximately 97% accuracy for detecting a single guardrail type and 86% accuracy for detecting multiple guardrail types.

It is critical to monitor the wearing of PPE on construction sites. PPE detection requires the detection of construction workers first and then identifying the presence of PPE on the detected workers. Fang et al. (2018) applied the Single Shot Detector (SSD) (Liu et al. 2016) to detect construction workers and used a CNN model for PPE wearing classification. Mneymneh et al. (2019) leveraged feature-based descriptors and classifiers to accomplish worker detection and monitoring of hardhat wearing.

CV-based methods were also applied in safety training to improve workers' safety awareness on sites. Jeelani et al. (2018) used eye-tracking techniques to obtain workers' eye-viewing patterns, and then integrated them with CV methods to localize their gaze position, which helps assess the ability to identify different hazards. Ding et al. (2018) applied a hybrid model of CNN and long short-term memory (LSTM) to identify construction activities that may cause potential hazards from videos. The identified activities can then be used as visual feedback to educate workers on how to perform activities safely.

Recent research has focused on analyzing the biomechanics of construction workers to proactively prevent musculoskeletal disorders (MSDs), which are work-related injuries among construction workers. Pose estimation in the CV domain has been commonly used for biomechanical analysis. A typical workflow of biomechanical analysis is that CV-based methods first extract 3D skeleton models of workers from visual data. The extracted skeleton model's 3D joint positions and joint angle features are then fed to pre-trained classifiers to detect ergonomic hazards. Han and Lee (2013) leveraged the HOG feature descriptor to extract 2D skeletons from video frames. Then, the 3D skeletons of workers were created from multiple 2D skeletons through triangulation. Kong et al. (2018) presented a framework to investigate the physical intensity of workers using a 3D biomechanical model. The model first estimated the 2D coordinates of each joint of construction workers using a deep neural network. Then another deep learning algorithm was applied to infer the 3D joint coordinates based on 2D joint coordinates. Lastly, the estimated 3D joint coordinates were combined with other collected data, such as foot pressure and plantar accelerations, to analyze the maximum mechanical energy expenditure for the whole body pose. Additional studies of CV-based methods applied in safety monitoring can be found in Table 1.

4.2 Productivity Improvement

Pose estimation and object detection domains were used to improve construction productivity. Pose estimation can help to improve construction productivity in two ways: first, position data from pose estimation of articulated construction equipment, such as excavators, can be used as designed profiles either to guide operators to complete tasks more efficiently or to guide equipment to finish tasks autonomously. Second, pose estimation can provide more comprehensive information for activity recognition, which can monitor construction activities in real time. Thus, it can provide visual feedback to managers and workers to understand the working status, and analyze the interactions between different construction entities.

TABLE. 1: Additional studies of CV-based methods applied to safety monitoring

CV Domains	Scenarios	Authors (Year)	Descriptions
Object Detection	Struck-by hazards monitoring	Zhang et al. (2020)	Leveraged Faster R-CNN to identify spatial relationships among construction entities.
		Yan et al. (2020)	Estimated the spatial proximity of construction entities based on the 3D bounding box reconstruction, and depth estimation from 2D monocular vision data.
		Luo et al. (2020)	Applied the YOLO-V2 (Redmon and Farhadi 2016) to monitor the proximity between workers and excavators.
		Guo et al. (2020)	Analyzed the rotation and orientation of construction vehicles from UAV images using a novel network.
Object Tracking	Struck-by hazards monitoring	Konstantinou and Brilakis (2018)	Matched the same construction workers in several camera views and used triangulation to generate 3D trajectories of workers.
		Hu et al. (2020)	Obtained the spatial information of construction machines and workers using background subtraction. A real-time safety value was computed based on the established safety rules to identify the safety status of the monitored objects.
		Cai and Cai (2020)	Applied Faster R-CNN to detect construction workers to obtain their trajectory in terms of 2D coordinates on an image plane.
		Tang et al. (2020)	Utilized a long short-term memory (LSTM) encoder-decoder followed by a mixture density network to forecast the motion trajectory of the construction equipment and workers for proactive safety analysis.
Object Detection	Fall Hazards Prevention	Fang et al. (2018)	Detected workers using a Faster R-CNN; applied a deep CNN model to the detected workers to identify the presence of a harness attached to workers.
	PPE Monitoring	Tang et al. (2020)	Proposed a human-object interaction (HOI) model to directly identify the work-tool interactions on construction sites for safety compliance checking.
Pose Estimation	Biomechanical Analysis	Zhang et al. (2018)	Leveraged a multi-stage CNN architecture to estimate the 3D location of different joints from video frames and created 3D skeletons of workers; the 3D joint position and joint angle features were extracted for ergonomic posture classification using machine learning techniques.
		Chu et al. (2020)	Assessed the ergonomic behavior of construction workers with the following steps: human body tracking from visual data; identifying and refining 2D joints from the tracking results; 3D reconstruction of the 2D joints; and the calculation of joint angles.

Researchers have been using different CV methods to extract the skeletons of articulated construction equipment. Marker-based pose estimation approaches were applied (Feng et al. 2018, Azar et al. 2015) to establish correspondences between 2D geometry features of excavators and camera pose in a certain coordination system.

After obtaining the 2D skeletons of excavators, Soltani et al. (2018) leveraged stereo vision to construct 3D coordinates of excavators' poses through triangulation. Xu and Yoon (2019) proposed a method to estimate the pose of the excavator manipulator for the feedback control system. Three neural networks were developed to estimate the displacements of the boom, arm, and bucket, respectively. The estimated displacements were then used as inputs to the Proportional Integral controller, which controls the manipulator model during the simulation process. Luo et al. (2020) leveraged three different deep neural networks to predict the key points of construction equipment from surveillance images. The proposed framework showed the potential to automatically capture construction equipment's pose and movement from surveillance videos. A hybrid model built with CNN and LSTM

(Kim and Chi 2020) was developed to recognize activities for earthmoving equipment, followed by an interaction analysis to analyze the productivity of the analyzed construction equipment.

Similar to the biomechanical analysis, researchers extracted the pose of construction workers and then classified the estimated pose data into different construction activities for automatic assessment of worker activities. Gong et al. (2011) and Khosrowpour et al. (2014) applied feature-based descriptors to extract and represent the skeletons of construction workers and then used classifiers to recognize specific activities. Luo et al. (2018) applied the improved three streams CNN and Temporal Segment Networks (TSNs) to recognize various construction activities. Roberts et al. (2020) leveraged YOLO-V3 (Redmon and Farhadi 2018) to detect construction workers from video frames and then used a pose estimator to estimate the body joints of each detected worker. The actual activity was determined by processing the detected pose with a proposed pose feature extraction module and reasoning the representation of pose-level spatiotemporal change within a short time window.

Object detection methods have also been used to identify and localize construction equipment and workers. Given the applied context reasoning and the identified working status of equipment and workers, the productivity rates of corresponding construction activities can be further analyzed. Spatial-temporal features combined with the SVM classifier were previously used for detecting earth-moving equipment (Golparvar-Fard et al. 2013). Markers were attached to earthmoving equipment to distinguish individual equipment (Rezazadeh Azar 2016). Recently, efforts have been made to leverage different R-CNNs to detect construction entities. Luo et al. (2018) proposed a method to identify diverse and concurrent activities executed by multiple objects in still site images. The Faster R-CNN was applied for object detection. Then rule-based semantic relevance and spatial relevance were introduced, together with 20 pre-defined activity patterns to recognize construction activities. Kim et al. (2018) applied the region-based fully convolutional networks (R-FCNs) to detect earthmoving entities and their locations from video frames. The detected results were then converted to the state and event information as inputs for productivity analysis. Fang et al. (2020) leveraged Mask R-CNN to segment different construction entities from images taken from a monocular camera and assign semantic information to the segmented instances. Then, a priori knowledge model was applied to estimate the location of detected entities, using the calculated vertical projections of entities in an aerial environment.

4.3 Progress Monitoring

Object detection and 3D reconstruction from CV domains were leveraged to facilitate the monitoring of construction progress. Object detection was focused on inspecting a specific construction task, while 3D reconstruction techniques aimed to create a larger field of view of construction sites to support the decision-making of project managers.

Hui et al. (2015) leveraged object detection methods to automatically count the number of bricks on a façade to monitor the progress of a bricklaying task. Czerniawski et al. (2016) proposed a pipe spool recognition method from point clouds to track pipe installation progress. The curvature-based shapes were first filtered using a shape descriptor from a 3D point cloud scene. Then the actual pipe objects were extracted using feature matching. For indoor construction, object detection methods were used to interpret the actual state of construction activities. Hamledari et al. (2017) presented four object detection modules, which automatically detected studs, insulation, electrical outlets, and drywalls. The current state of the interior partition installation can be inferred based on the presence of the above-mentioned detected objects. Kropp et al. (2018) registered indoor images with 4D Building Information Modeling (BIM) at the same timestamp on the construction schedule and then detected the presence of specific indoor objects to monitor the progress of construction activities. Deng et al. (2020) leveraged the handcrafted descriptor methods to detect tiles and compared them with the as-designed model (BIM). Edge detection methods were also used to identify the boundaries of the tiles for the area quantification. The transformation relationship between a real-world coordinate system and an image coordinate system can be established by a camera calibration algorithm. Braun et al. (2020) proposed an image-based object detection method using the Mask R-CNN to detect formwork elements from construction images. The detection results will then be compared with as-planned BIM to identify the status of certain construction progress. Similarly, Pour Rahimian et al. (2020) leveraged the semantic segmentation algorithm called FuseNet (Hazirbas et al. 2017) to recognize construction elements from RGB-D images and superimposed the detected elements with BIM in a gamelike immersive virtual environment.

3D point clouds of a construction site have been created using images captured on sites to represent the as-built conditions of projects. Kim et al. (2018) leveraged CV-based methods to detect common features between images and automatically merge point clouds with overlapped areas to improve the point clouds creation process. The generated point clouds can then be integrated with BIM in the same coordination system to identify deviations between the as-designed and the as-built conditions (Golparvar-Fard Mani et al. 2015). The identified deviations can be used as supporting information for decision-making in terms of a real-time schedule update. Han et al. (2018) integrated both object detection and 3D reconstruction. The BIM was first used as a geometry constraint to filter out the non-relevant parts of point clouds. The material type of specific building elements can be identified through object detection methods for operation-level activities progress monitoring.

4.4 Infrastructure Inspection

Researchers have been exploring the benefits of using CV-based methods during the asset management phase, especially for inspecting the structural health of infrastructures. Aging, degradation, and deterioration will cause the potential failure of structural systems. Thus, constant structural health monitoring is critical to maintaining the durability and integrity of structural systems. Commonly applied methods include image processing, 3D reconstruction, and object detection.

Image processing techniques were commonly applied to reduce background noise in visual data, which improves the performance of extracting features of certain defects. Li et al. (2016) applied image processing methods to detect the location and shape of potholes from 2D images. The image smoothing technique was first applied to reduce the variance of the pixel density of the pavement surface without significantly affecting the defect information. Uniform background illumination can also affect the detection of potholes in images. The image normalization technique (Oliveira and Correia 2014) was used to normalize the pixel intensities of non-defect areas while maintaining the pixel intensities of the defect areas. The images were then further processed by implementing an intensity saturation procedure. The enhanced images were fused with 3D radar data to analyze the positions and dimensions of potholes. Image processing methods were also used to enhance the features of interest. Davoudi et al. (2018) applied Canny's edge detection method (Canny 1986) to segment concrete crack patterns from the image background. The morphological transformation was then used to make all cracks have a consistent single-pixel width. The processed images were then fed into a supervised machine learning model to quantitatively analyze the damage levels for future load prediction.

3D reconstruction techniques allow a larger field of view via reconstructing collected visual data, which is critical for the structural inspection of large-scale infrastructure projects. Studies (Siebert and Teizer 2014, Rodriguez-Gonzalvez et al. 2014) showed that image-based 3D reconstruction from UAV images of large-scale and complex infrastructure scenarios could capture the as-built condition in a more quick and affordable way, and eliminate the limitation of laser scanners and other classical topographical survey methods, including accessibility and safety issues. Golparvar-Fard Mani et al. (2015) leveraged an image-based 3D reconstruction pipeline consisting of Structure from Motion (SfM) and multi-view stereo (MVS) (Golparvar-Fard Mani et al. 2015) to create 3D point clouds of highway assets from video frames. The developed 3D point clouds can overcome the issues of unstable illumination and resolution conditions, and static and dynamic occlusions from 2D visual data. The created 3D point clouds can also serve as a platform to geo-register the 2D image for asset segmentation and object detection. Besides 3D reconstruction, Bang et al. (2017) proposed an image stitching method to create a panorama view of a large-scale construction site from UAV images. The generated panorama view can also help project managers better understand the as-built conditions of a large-scale and complex construction site.

Object detection has been used for detecting the type, location, and measuring the size of structural defects from visual data. Different features distinguished different types of defects. Features can be extracted by manually created feature descriptors. Halfawy and Hengmeechai (2014) applied the HOG descriptor to identify the features of sewer defects. The identified vectors were then fed into the SVM classifier to determine the presence of defects. More researchers have been using various deep neural networks in the infrastructure inspection areas for the past three years. Cheng and Wang (2018) leveraged the Faster R-CNN to detect the location of sewage pipe defects from surveillance video frames. After extracting feature maps from images using CNN, Faster R-CNN leverages Regional Proposal Network (RPN) to generate regions of interest (RoIs) based on the extracted feature maps. The RoIs were then used as input for the detector to create bounding boxes around the detected defects. Pan et al. (2020) used a novel deep neural network to automatically detect concrete cracks at a pixel level. The backbone of the novel network was based on VGG19. This novel network achieved better performance than the Fully

Convolutional Neural Network (FCN) applied by Dung and Anh (2019) because it used the self-attention mechanism. Specifically, the self-attention mechanism reduced computational cost by focusing on RoIs and encoding rich contextual information into extracted local features. The proposed model increased Mean Intersection over Union (IoU) to 85.31%. More studies of CV-based methods applied in infrastructure inspection can be found in Table 2.

TABLE. 2: Additional studies of CV-based methods applied to infrastructure inspection

CV Domains	Scenarios	Authors (Year)	Descriptions
Image Processing	Sewage Inspection	Halfawy and Hengmeechai (2014b)	Leveraged the optical flow technique to estimate the camera motion parameter from sewer inspection videos; the change of camera motion will indicate the presence of defects in a sewer pipe.
	Potholes Identification	Ouma and Hahn (2017)	Applied background noise reduction and image smoothing techniques, followed by fuzzy c-means clustering and morphological reconstruction to extract potholes.
	Spalling Detection	Dawood et al. (2017)	Applied image smoothing method to reduce the background noise; used a Gaussian Blur filter to enhance the image resolution; used several color filters and image scaling to reveal the difference in the intensity level of the defect areas.
	Structural Displacements Tracking	Havaran and Mahmoudi (2020)	Applied grayscale and edge detection techniques to images from video frames; used the Randomized Hough transform technique (Basca et al. 2005) to detect the ellipse markers' movement in order to determine the structural displacement.
3D Reconstruction	Bridge Inspection	Morgenthal et al. (2019)	Used the generated 3D point clouds to perform structural analysis and geometrical change detection since they can provide information about the current deformation state and enable a mechanical interpretation of the acquired data.
		Xie et al. (2018)	Leveraged the image stitching technique to create a panorama of a concrete bridge bottom surface using inspection images.
Object Detection	Bridge Inspection	Asadi et al. (2020)	Used a handcrafted HOG feature descriptor to detect the coordination of rebars from the concrete bridge deck Ground Penetrating Radar (GPR) images; the GPR images were used for concrete bridge deck analysis, including deterioration levels analysis, repair cost prediction, underground utility tracing.
		Deng et al. (2020)	Developed a novel deep neural network to detect structural damage of bridges from image data. The network achieved the mean IoU at 61.95%.
	Crack Detection	Kong and Li (2019)	Proposed a feature descriptor to extract image features that were used to align images taken from different camera poses in the same coordinate system to detect fatigue cracks.
	Ceiling Panels Inspection	Wang et al. (2020)	Leveraged a 15 layers CNN model to detect and localize the damaged ceiling panels in large-span structures
	Pipe Crack Detection	Zuo et al. (2020)	Used the structured edge detection method (Dollár and Zitnick 2015) to detect the cracks of pipes from inspection videos.
Object Classification	Bridge Inspection	Shen et al. (2018)	Leveraged a three-layer artificial neural network model to classify the intensity and severity of rust of steel bridges from images.
	Sewage Inspection	Kumar et al. (2018)	Leveraged the CNN model to classify different sewage defects from root intrusions, deposits, and cracks.

4.5 Robotic Application

Over the past few years, there has been a dramatic increase in robotic research in construction fields to reduce the costs and safety risks and automate part of the construction process. CV plays an important role in the construction robotics field. In general, CV applications in construction robotics can be divided into two fields: path planning and onboard inspection (Lattanzi and Miller 2017).

Path planning is critical for construction robots to minimize traveling time while achieving maximum coverage and avoiding obstacles. Asadi et al. (2018) presented a vision system implemented on an unmanned ground vehicle (UGV). The proposed system contained the monocular Simultaneous Localization and Mapping (SLAM) module and contextual scene understanding algorithms. The SLAM module constantly mapped the surrounding environment, and contextual scene understanding algorithms performed pixel-wise segmentation for proper context-aware motion planning. The integrated robotic system achieved real-time running performance and demonstrated the feasibility of deploying an autonomous robotic system in the future. Path planning methods were also applied to unmanned aerial vehicles (UAVs). Bolourian and Hammad (2020) presented a path planning method for LiDAR-equipped UAVs to perform bridge inspections. The method achieved obstacle avoidance and coverage calculation, resulting in minimum travel and processing time. The method was based on a pre-developed 3D bridge model. The bridge surface was divided into cells for detailed visibility analysis. The Importance Values (IVs) were assigned to each cell based on the level of criticality of the bridge surface. The level of criticality was determined by the depth and angle of surface cracks, which were analyzed using LiDAR ray tracing. A set of View Points of Interest (VPIs) were developed based on the assigned IVs, and the optimum path was generated based on VPIs using the Genetic Algorithm.

Asadi et al. (2020) proposed a novel cooperative platform to integrate UGV and UAV for construction inspection or monitoring. The main function of this cooperative platform was to navigate UGV and UAV in the construction site, collect visual data and map the space. The UGV will scan the space of interest assigned by construction teams by navigating itself around the construction site while the UAV follows. If a space is not accessible to the UGV, the UGV will send a UAV to scan that space. The UAV will then return and follow the UGV to the next space of interest. SLAM and image segmentation techniques were adopted in the proposed platform. The SLAM can help unmanned vehicles determine their positions while building up a map of the scanned space. The image segmentation can help unmanned vehicles understand the scene better by pixel-wise labeling images.

Many inspection tasks of construction projects require data gathering from unreachable spaces or tightly enclosed areas. With the versatility of robotic systems, robots for inspection tasks can ease the burden on field inspectors by having access to the challenging areas and navigating themselves (semi-) autonomously while capturing data from the sites. Menendez et al. (2018) presented a robotic system used for tunnel structural inspection. The vision system of robotic system was composed of two pairs of cameras and an onboard lighting system. The first pair of cameras were used to detect various concrete defects in tunnels, such as spalling, efflorescence, and cracks. The detection method was a pre-trained CNN model. The second pair of cameras were used to extract 3D information on defects, such as positions and orientations, which were used to guide robotic cranes and arms to the corresponding place. Similarly, McLaughlin et al. (2020) presented a robotic platform to detect defects in concrete bridges. The vision system of the robotic platform contained a camera and a LiDAR scanner. By fusing the LiDAR data with the CNN-labeled image data, semantically labeled 3D point clouds were generated, which can be used to quantify the defect sizes. The presented process provided a much safer and more efficient way than normal bridge inspection routines. More studies of CV-based methods in robotic applications can be found in Table 3.

TABLE. 3: Additional studies of CV-based methods applied to robotic application

CV Domains	Scenarios	Authors (Year)	Descriptions
Object Detection	Path Planning	Yoder and Scherer (2016)	The planning algorithm leveraged two fisheye cameras as visual inputs. For each given view, different cells of the visible surface were weighted by their utility based on vehicle location, partial map, and exploration views. The navigation path was generated based on either if the view was reachable or the level of view utility.
Pose Estimation/ 3D Reconstruction	Path Planning	Feng et al. (2015)	Used planar markers to help a UGV to localize itself; a single camera was used to detect the visual marker to establish local reference frames. 3D point clouds were generated by a mounted laser scanner, and were then registered into the same reference frame established by the marker-based method.
Object Detection/ Object Tracking	Path Planning	Kim et al. (2020)	Adopted the YOLO-V3 network to monitor targets' locations and trajectories, and the Social-GAN (Gupta et al. 2018) network to predict targets' future trajectories and proximity.

CV Domains	Scenarios	Authors (Year)	Descriptions
Object Detection	Data Gathering	Atkinson et al. (2020)	Used a UGV to identify different regions of the underfloor scene, such as floorboards, joists, air vents, and pipework using the Mask R-CNN algorithm.
		Wang et al. (2019)	Developed a robot that recycles wasted nails and screws while inspecting the construction site. The robot leveraged Faster R-CNN to detect and localize objects on the ground.

5. DISCUSSIONS

A trending timeline was developed for each reviewed paper to identify the changing popularity of construction use cases and CV domains within the analyzed period (see Fig. 2). In general, the total number of papers that focus on CV applications in the construction industry each year has increased. The variety of use cases and CV domains has also increased over time. In the early 2010s, studies covered only two or three construction use cases per year, and there was no paper reviewed focusing on robotic applications in construction. In the last five years, studies have covered almost all categorized use cases and CV domains. It is worth noticing that there were overlaps between CV domains and construction use cases. Each reviewed paper did not necessarily focus on just one construction use case, or implement one single CV domain. Multiple CV-based methods could be applied in one construction use case, and a single CV-based method could also be implemented to fulfill tasks in various construction use cases. In this case, each CV domain and each use case was counted within the year, which resulted in the total number of construction use cases and CV domains exceeding the number of reviewed papers in some years.

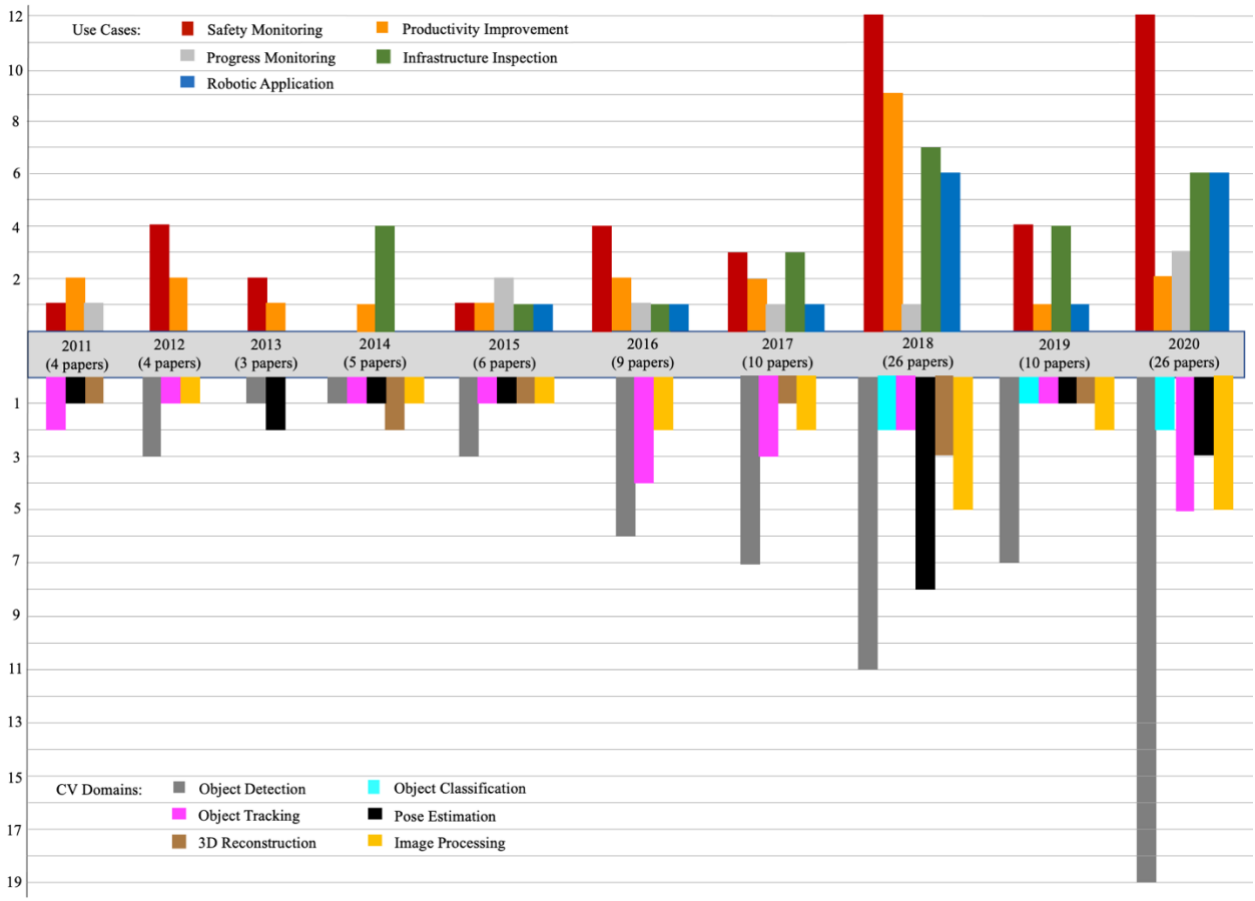

FIG. 2: A trending timeline for each reviewed paper within the analyzed time frame

Safety monitoring was the use case that gained the most attention from researchers (43 papers), followed by infrastructure inspection (26 papers) and productivity improvement (23 papers). One interesting observation was that robotic applications in the construction area have started to gain popularity among researchers in recent years. Within CV, the most frequently applied methods were included in the object detection domain (58 papers). Object tracking (20 papers) and image processing (19 papers) were also very popular domains that studies tended to use. The reason for the gaining popularity of both construction use cases and CV domains is the introduction of deep learning models to the construction industry, and the advancement of computing powers and visual data acquisition platforms. Most deep learning models focus on object detection, which can be applied to various construction use cases.

The CV domain that was applied the least was object classification. Construction scenes were rather complex, with a lot of moving entities. Only producing a list of object categories presented in the visual data without providing the position and scale information was the most significant limitation of implementing object classification in construction and asset management processes. The least popular construction use case was progress monitoring. This can be one of the directions for future research regarding CV-based methods application in the construction industry.

While CV-based approaches bring many benefits to construction and asset management processes, some remaining challenges and limitations need to be considered by companies willing to adopt CV in their construction tasks. The identified implementation challenges and technical limitations were summarized from the reviewed papers.

5.1 Implementation Challenges

Despite the successful implementation of CV techniques in the construction industry, some challenges persist in applying CV-based methods in construction and asset management processes. Some of the challenges were observed from the reviewed literature, including (1) lack of interdisciplinary knowledge, (2) data privacy protection, and (3) cost.

5.1.1 Lack of Interdisciplinary Knowledge

Finding construction engineers with sufficient CV knowledge to apply CV-based methods to different construction tasks can be challenging. Most engineers with construction backgrounds are not familiar with CV/deep learning. Similarly, experts from CV fields generally do not have an understanding of basic construction processes. Although knowledge from both fields is acquirable, CV knowledge usually has a steep learning curve, while construction expertise requires onsite experience. People with interdisciplinary knowledge and who can bring CV-based solutions to different construction use cases are scarce (Fang et al. 2020).

5.1.2 Data Privacy Protection

CV-based approaches need a large amount of data to support their successful implementation. Images and videos used for safety and productivity improvement purposes may contain the faces and activities of construction workers. Pose estimation also needs pose data from workers. It is critical to have regulations and standards that identify how to use personal data properly, and protect personal data privacy. Companies doing business in the European Union (EU) must comply with the EU's General Data Protection Regulation (GDPR), which requires companies to provide meaningful information about personal data usage and provides individuals the right to remove their personal data. Deep learning researchers are also making efforts to introduce guides as to what level of explanation is needed for an algorithm to prevent ethical complications (Akinosho et al. 2020). In the future, construction companies still need to face this challenge, and be cautious about collecting personal data on sites for deep learning usage.

5.1.3 Cost

The implementation of state-of-the-art CV-based methods will add extra cost to construction projects. Although CV-based methods aim to lower the overall cost by improving productivity and reducing safety hazards, the front-end cost of technology implementation is inevitable. Leveraging deep learning models requires machines with powerful computational capabilities. Companies willing to adopt deep learning techniques need to either purchase machines, or pay for cloud computing subscriptions. Technology implementation usually also comes with consulting service, which is another cost besides the hardware part. Construction companies should find cost-efficient approaches to adopt deep learning techniques in the industry (Akinosho et al. 2020).

5.2 Technical Limitations

Besides the implementation challenges mentioned above, deep learning, or CV in general, remains some technical limitations while being implemented in the construction industry. Some of the limitations were identified from the reviewed literature, including (1) lack of training data; (2) Non-transparency; and (3) compatibility with other data sources.

5.2.1 Lack of Training Data

Successful identification and detection of deep learning models require a large number of datasets for training. Adequate training data can ensure the inter and intra-class variability of the trained model, while limited training data can yield low performance or overfitting. Compared with some benchmark image datasets (e.g., ImageNet, Open Images, and COCO), which contain numerous labeled training images, the collected construction images need to be manually labeled and tagged by researchers. This can be a tedious and time-consuming process.

Solutions to address the limited data issue are transfer learning techniques and data augmentation. The basic idea of transfer learning is to select a deep neural network pre-trained with large image datasets as a starting point. Then the selected network is tuned with additional images containing objects for tasks of interest. Kim et al. (2018) used the pre-trained ResNet-50 model using the ImageNet dataset and fine-tuned it with construction vehicle datasets for transfer learning. Similarly, Kim et al. (2019) used the pre-trained YOLO-V3 model from ImageNet, and fine-tuned it with image data involving a wide range of construction entities, enhancing the object detection capability. Data augmentation is to generate synthetic images by projecting either CAD models or reconstructed 3D models (Kim and Kim 2018) from different points of view. This approach can significantly increase the quantity of training data. Various views of a construction excavator CAD model (Soltani et al. 2017) and a railing CAD model (Kolar et al. 2018) were created for training the corresponding object detectors. Kim and Kim (2018) used a reconstructed 3D dump truck model to generate synthetic images. Other data augmentation approaches include rotation, shifts, and flips of the existing image dataset to enhance the overall amount of the training dataset.

5.2.2 Non-transparency of Deep Learning Models

Deep learning models usually have a non-linear multilayer structure, and their predictions are not traceable by humans (Buhrmester et al. 2019), and thus are criticized for being non-transparent. Although some strides have been made in visualizing the contributions of individual nodes in complex networks, the transparency issue has not yet been solved (Marcus 2018). Researchers are not able to determine the exact features extracted from nodes and are not able to identify the weights or parameters that need to adjust for each layer to achieve better performance. The opacity of deep learning models leads to a potential liability of applying deep learing models, especially in construction safety scenarios.

5.2.3 The Integration with Other Data Sources

While CV-based methods can interpret visual data for ongoing operations, it is critical to realize that the analyzed visual information needs to be integrated with other devices or data sources to complete a broader range of construction tasks. For the safety monitoring use case, safety rules are constantly changing based on different scenarios or processes. Therefore, CV-based methods need to be adjusted accordingly to accommodate the corresponding safety rules better (Fang et al. 2020). CV-based methods can also be aided by other devices (e.g., sensors, RFID), which are capable of acquiring other types of data. The acquired data can be integrated with the information analyzed from CV-based methods to generate a more comprehensive framework to facilitate more complex use cases. For example, accelerometer sensors combined with CV techniques can keep track of construction entities under occlusion.

6. CONCLUSION AND FUTURE WORK

A systematic literature review was conducted in this paper to summarize the CV-based methods applied to various construction and asset management processes. A framework was developed to categorize different construction use cases and CV domains from the reviewed journal papers. In total, five use cases and six domains were generated. A trending timeline was developed over the reviewed time frame. The trending timeline showed the changing amount of journal papers published in the identified research area and the trending popularity in both construction use cases and CV domains. The most popular construction use cases were safety monitoring, followed by infrastructure inspection and productivity improvement. Robotic applications have been gaining more attention

in recent years. The most implemented CV domain was object detection, followed by object tracking and image processing. Even though CV-based approaches are beneficial to the construction industry, there are still many implementation challenges and technical limitations that companies and researchers should be aware of.

An emerging area of research is using computer vision to support construction robotics on jobsites. Specifically related to computer vision, there is a need to pursue the following items: generating large and comprehensive image datasets specifically for the construction industry; understanding the decision-making process of deep learning models, so that people can tune models based on their needs; and applying different CV domains (pose estimation, 3D reconstruction) to support construction robotic applications. In addition, one particular topic of interest for using visual data from construction sites is to develop standards and practices for collecting and protecting the personal data of construction workers.

REFERENCES

Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O., and Ahmed, A. A. (2020). "Deep learning in the construction industry: A review of present status and future innovations. " *Journal of Building Engineering*, 32, 101827.

Asadi, K., Kalkunte Suresh, A., Ender, A., Gotad, S., Maniyar, S., Anand, S., Noghabaei, M., Han, K., Lobaton, E. and Wu, T. (2020). "An integrated UGV-UAV system for construction site data collection." *Automation in Construction*, 112, 103068.

Asadi, K., Ramshankar, H., Pullagurla, H., Bhandare, A., Shanbhag, S., Mehta, P., Kundu, S., Han, K., Lobaton, E. and Wu, T. (2018). "Vision-based integrated mobile robotic system for real-time applications in construction." *Automation in Construction*, 96, 470–482.

Asadi, P., Gindy, M., Alvarez, M. and Asadi, A. (2020). "A computer vision based rebar detection chain for automatic processing of concrete bridge deck GPR data." *Automation in Construction*, 112, 103106.

Atkinson, G. A., Zhang, W., Hansen, M. F., Holloway, M. L. and Napier, A. A. (2020). "Image segmentation of underfloor scenes using a mask regions convolutional neural network with two-stage transfer learning." *Automation in Construction*, 113, 103118.

Azar, E. R., Feng, C. and Kamat, V. R. (2015). "Feasibility of in-plane articulation monitoring of excavator arm using planar marker tracking." *Journal of Information Technology in Construction (ITcon)*, 20(15), 213–229.

Babatunde, O. H., Armstrong, L., Leng, J. and Diepeveen, D. (2015). "A survey of computer-based vision systems for automatic identification of plant species." *Journal of Agricultural Informatics*, Hungarian Association of Agricultural Informatics European Federation for Information Technology in Agriculture, Food and the Environment, 6(1), 61–71.

Bang, S., Kim, H. and Kim, H. (2017). "UAV-based automatic generation of high-resolution panorama at a construction site with a focus on preprocessing for image stitching." *Automation in Construction*, 84, 70–80.

Basca, C. A., Talos, M. and Brad, R. (2005). "Randomized hough transform for ellipse detection with result clustering." *EUROCON 2005 - The International Conference on "Computer as a Tool,"* 1397–1400.

Best, R. D. (2021). Statista, < https://www.statista.com/statistics/192049/value-added-by-us-construction-as-a-percentage-of-gdp-since-2007 > (accessed 27 April 2021).

Bolourian, N. and Hammad, A. (2020). "LiDAR-equipped UAV path planning considering potential locations of defects for bridge inspection." *Automation in Construction*, 117, 103250.

Bora, K., Chowdhury, M., Mahanta, L. B., Kundu, M. K. and Das, A. K. (2016). "Pap smear image classification using convolutional neural network." *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing - ICVGIP '16*, ACM Press, Guwahati, Assam, India, 1–8.

Braun, A., S. Tuttas, A. Borrmann, and U. Stilla. 2020. “Improving progress monitoring by fusing point clouds, semantic data and computer vision.” *Automation in Construction*, 116: 103210. https://doi.org/10.1016/j.autcon.2020.103210.

Brilakis, I., Park, M.-W. and Jog, G. (2011). "Automated vision tracking of project related entities." *Advanced Engineering Informatics*, Special Section: Advances and Challenges in Computing in Civil and Building Engineering, 25(4), 713–724.

Buhrmester, V., Münch, D. and Arens, M. (2019). "Analysis of explainers of black box deep neural networks for computer vision: A survey." *arXiv:1911.12116 [cs]*.

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." *Journal of Computing in Civil Engineering*, 34(4), 04020021.

Canny, J. (1986). "A computational approach to edge detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6), 679–698.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." *arXiv:1812.08008 [cs]*.

Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y. (2017). "Realtime multi-person 2D pose estimation using part affinity fields." *arXiv:1611.08050 [cs]*.

Chan, T., Jia, K., Gao, S., Lu, J., Zeng, Z. and Ma, Y. (2015). "PCANet: A simple deep learning baseline for image classification?" *IEEE Transactions on Image Processing*, 24(12), 5017–5032.

Chen, J., Fang, Y. and Cho, Y. K. (2017). "Real-time 3D crane workspace update using a hybrid visualization approach." *Journal of Computing in Civil Engineering*, 31(5), 04017049.

Cheng, J. C. P. and Wang, M. (2018). "Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques." *Automation in Construction*, 95, 155–171.

Chu, W., Han, S., Luo, X. and Zhu, Z. (2020). "Monocular vision–based framework for biomechanical analysis or ergonomic posture assessment in modular construction." *Journal of Computing in Civil Engineering*, 34(4), 04020018.

Comaniciu, D., Ramesh, V. and Meer, P. (2003). "Kernel-based object tracking." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5), 564–577.

Cristianini, N. and Shawe-Taylor, J. (2000). "An introduction to support vector machines and other kernel-based learning methods". *Cambridge University Press*, Cambridge.

Czerniawski, T., Nahangi, M., Haas, C. and Walbridge, S. (2016). "Pipe spool recognition in cluttered point clouds using a curvature-based shape descriptor." *Automation in Construction*, 71, 346–358.

Dalal, N. and Triggs, B. (2005). "Histograms of oriented gradients for human detection." *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 886–893 vol. 1.

Davoudi, R., Miller, G. R. and Kutz, J. N. (2018). "Data-driven vision-based inspection for reinforced concrete beams and slabs: Quantitative damage and load estimation." *Automation in Construction*, 96, 292–309.

Dawood, T., Zhu, Z. and Zayed, T. (2017). "Machine vision-based model for spalling detection and quantification in subway networks." *Automation in Construction*, 81, 149–160.

Deng, H., Hong, H., Luo, D., Deng, Y. and Su, C. (2020). "Automatic indoor construction process monitoring for tiles based on BIM and computer vision." *Journal of Construction Engineering and Management*, 146(1), 04019095.

Deng, W., Mou, Y., Kashiwa, T., Escalera, S., Nagai, K., Nakayama, K., Matsuo, Y. and Prendinger, H. (2020). "Vision based pixel-level bridge structural damage detection using a link ASPP network." *Automation in Construction*, 110, 102973.

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." *Automation in Construction*, 86, 118–124.

Dollár, P. and Zitnick, C. L. (2015). "Fast edge detection using structured forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8), 1558–1570.

Dung, C. V. and Anh, L. D. (2019). "Autonomous concrete crack detection using deep fully convolutional neural network." *Automation in Construction*, 99, 52–58.

Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2018). "RMPE: Regional Multi-person Pose Estimation." *arXiv:1612.00137 [cs]*.

Fang, Q., Li, H., Luo, X., Ding, L., Luo, H. and Li, C. (2018). "Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment." *Automation in Construction*, 93, 148–164.

Fang, W., Ding, L., Luo, H. and Love, P. E. D. (2018). "Falls from heights: A computer vision-based approach for safety harness detection." *Automation in Construction*, 91, 53–61.

Fang, W., Ding, L., Zhong, B., Love, P. E. D. and Luo, H. (2018). "Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach." *Advanced Engineering Informatics*, 37, 139–149.

Fang, W., Love, P. E. D., Luo, H. and Ding, L. (2020). "Computer vision for behaviour-based safety in construction: A review and future directions. " *Advanced Engineering Informatics*, 43, 100980.

Fang, W., Zhong, B., Zhao, N., Love, P. E. D., Luo, H., Xue, J. and Xu, S. (2019). "A deep learning-based approach for mitigating falls from height with computer vision: Convolutional neural network." *Advanced Engineering Informatics*, 39, 170–177.

Fang, Q., Li, H., Luo, X., Li, C. and An, W. (2020). "A sematic and prior-knowledge-aided monocular localization method for construction-related entities." *Computer-Aided Civil and Infrastructure Engineering*, 35(9), 979–996.

Feng, C., Kamat, V. R. and Cai, H. (2018). "Camera marker networks for articulated machine pose estimation." *Automation in Construction*, 96, 148–160.

Feng, C., Xiao, Y., Willette, A., McGee, W. and Kamat, V. R. (2015). "Vision guided autonomous robotic assembly and as-built scanning on unstructured construction sites." *Automation in Construction*, 59, 128–138.

Feng, M., Gilani, S. Z., Wang, Y., and Mian, A. (2018). "3D Face Reconstruction from Light Field Images: A Model-free Approach." *ECCV 2018*, pp. 501–518.

Goesele, M., Snavely, N., Curless, B., Hoppe, H. and Seitz, S. M. (2007). "Multi-view stereo for community photo collections." *2007 IEEE 11th International Conference on Computer Vision*, IEEE, Rio de Janeiro, Brazil, 1–8.

Golparvar-Fard, M., Balali, V. and de la Garza, J. M. (2015). "Segmentation and recognition of highway assets using image-based 3D point clouds and semantic texton forests." *Journal of Computing in Civil Engineering*, 29(1), 04014023.

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." *Advanced Engineering Informatics*, 27(4), 652–663.

Golparvar-Fard, M., Peña-Mora, F. and Savarese, S. (2015). "Automated progress monitoring using unordered daily construction photographs and IFC-based building information models." *Journal of Computing in Civil Engineering*, 29(1), 04014025.

Gonzalez, R. C. and Woods, R. E. (2007). *Digital Image Processing*. Pearson, Upper Saddle River, N.J, ISBN 978-0-13-335672-4.

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation." *arXiv:1311.2524 [cs]*.

Girshick, R. (2015). "Fast R-CNN." *arXiv:1504.08083 [cs]*.

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." *Automation in Construction*, 20(8), 1211–1226.

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." *Advanced Engineering Informatics*, Special Section: Advances and Challenges in Computing in Civil and Building Engineering, 25(4), 771–782.

Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." *IEE Proceedings F (Radar and Signal Processing)*, IET Digital Library, 140(2), 107–113.

Guo, Y., Xu, Y. and Li, S. (2020). "Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network." *Automation in Construction*, 112, 103124.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S. and Alahi, A. (2018). "Social GAN: Socially acceptable trajectories with generative adversarial networks." *arXiv:1803.10892 [cs]*.

Halfawy, M. R. and Hengmeechai, J. (2014). "Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine." *Automation in Construction*, 38, 1–13.

Halfawy, M. R. and Hengmeechai, J. (2014b). "Optical flow techniques for estimation of camera motion parameters in sewer closed circuit television inspection videos." *Automation in Construction*, 38, 39–45.

Hamledari, H., McCabe, B. and Davari, S. (2017). "Automated computer vision-based detection of components of under-construction indoor partitions." *Automation in Construction*, 74, 78–94.

Han, K., Degol, J. and Golparvar-Fard, M. (2018). "Geometry- and appearance-based reasoning of construction progress monitoring." *Journal of Construction Engineering and Management*, 144(2), 04017110.

Han, S. and Lee, S. (2013). "A vision-based motion capture and recognition framework for behavior-based safety management." *Automation in Construction*, 35, 131–141.

Han, X.-F., Laga, H., and Bennamoun, M. (2021). "Image-based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5), 1578–1604.

Havaran, A. and Mahmoudi, M. (2020). "Markers tracking and extracting structural vibration utilizing Randomized Hough transform." *Automation in Construction*, 116, 103235.

Hazirbas, C., Ma, L., Domokos, C. and Cremers, D. (2017). "FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture." *Computer Vision – ACCV 2016*, Lecture Notes in Computer Science, S.-H. Lai, V. Lepetit, K. Nishino and Y. Sato, eds., Springer International Publishing, Cham, 213–228.

He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2018). "Mask R-CNN." *arXiv:1703.06870 [cs]*.

Hu, Q., Bai, Y., He, L., Cai, Q., Tang, S., Ma, G., Tan, J. and Liang, B. (2020). "Intelligent framework for worker-machine safety assessment." *Journal of Construction Engineering and Management*, 146(5), 04020045.

Huang, T. (1996). "Computer vision: Evolution and promise". *19th CERN School of Computing*, CERN, Geneva, 1996, pp. 21–25 https://doi.org/10.5170/CERN-1996- 008.21; ISBN 978-9290830955.

Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., LeGendre, C., Luo, L., Ma, C., and Li, H. (2018). "Deep Volumetric Video From Very Sparse Multi-view Performance Capture." *ECCV 2018*, pp. 336–354.

Hui, L., Park, M.-W. and Brilakis, I. (2015). "Automated brick counting for façade construction progress estimation." *Journal of Computing in Civil Engineering*, 29(6), 04014091.

Izadinia, H., Shan, Q., and Seitz, S. M. (2017). “IM2CAD.” *IEEE CVPR 2017*, pp. 5134–5143.

James, M. R. and Robson, S. (2012). "Straightforward reconstruction of 3D surfaces and topography with a camera: Accuracy and geoscience application." *Journal of Geophysical Research: Earth Surface*, 117(F3).

Jeelani, I., Han, K. and Albert, A. (2018). "Automating and scaling personalized safety training using eye-tracking data." *Automation in Construction*, 93, 63–77.

Jiang, Z., Messner, J. I. and Dubler, C. R. (2017). "Defining a taxonomy for virtual 3D city model use cases with a focus on facility asset management—a virtual campus case study." *Computing in Civil Engineering 2017*, American Society of Civil Engineers, Seattle, Washington, 43–50.

Jiang, Z. and Messner, J. I. (2020). “Computer vision-based methods applied to construction processes: A literature review.” *American Society of Civil Engineers*, 1233–1241.

Kalman, R. E. (1960). "A new approach to linear filtering and prediction problems." *Journal of Basic Engineering*, 82(1), 35–45.

Khanday, N. Y., and Sofi, S. A. (2021). “Taxonomy, state-of-the-art, challenges and applications of visual understanding: A review.” *Computer Science Review*, 40, 100374.

Khosrowpour, A., Niebles, J. C. and Golparvar-Fard, M. (2014). "Vision-based workface assessment using depth images for activity analysis of interior construction operations." *Automation in Construction*, 48, 74–87.

Kim, D., Lee, S. and Kamat, V. R. (2020). "Proximity prediction of mobile objects to prevent contact-driven accidents in co-robotic construction." *Journal of Computing in Civil Engineering*, 34(4), 04020022.

Kim, D., Liu, M., Lee, S. and Kamat, V. R. (2019). "Remote proximity monitoring between mobile construction resources using camera-mounted UAVs." *Automation in Construction*, 99, 168–182.

Kim, H., Kim, K. and Kim H. (2016). "Vision-based object-centric safety assessment using fuzzy inference: monitoring Struck-By Accidents with Moving Objects." *Journal of Computing in Civil Engineering*, 30(4), 04015075.

Kim, H. and Kim, H. (2018). "3D reconstruction of a concrete mixer truck for training object detectors." *Automation in Construction*, 88, 23–30.

Kim, H., Kim, H., Hong, Y. W. and Byun, H. (2018). "Detecting construction equipment using a region-based fully convolutional network and transfer learning." *Journal of Computing in Civil Engineering*, 32(2), 04017082.

Kim, J. and Chi, S. (2017). "Adaptive detector and tracker on construction sites using functional integration and online learning." *Journal of Computing in Civil Engineering*, 31(5), 04017026.

Kim, J. and Chi, S. (2020). "Multi-camera vision-based productivity monitoring of earthmoving operations." *Automation in Construction*, 112, 103121.

Kim, K., Kim, H. and Kim, H. (2017). "Image-based construction hazard avoidance system using augmented reality in wearable device." *Automation in Construction*, 83, 390–403.

Kim, P., Chen, J. and Cho, Y. K. (2018). "Automated point cloud registration using visual and planar features for construction environments." *Journal of Computing in Civil Engineering*, 32(2), 04017076.

Kitsikidis, A., Dimitropoulos, K., Douka, S. and Grammalidis, N. (2014). "Dance analysis using multiple Kinect sensors." *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, 789–795.

Kolar, Z., Chen, H. and Luo, X. (2018). "Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images." *Automation in Construction*, 89, 58–70.

Kong, L., Li, H., Yu, Y., Luo, H., Skitmore, M. and Antwi-Afari, M. F. (2018). "Quantifying the physical intensity of construction workers, a mechanical energy approach." *Advanced Engineering Informatics*, 38, 404–419.

Kong, X. and Li, J. (2019). "Non-contact fatigue crack detection in civil infrastructure through image overlapping and crack breathing sensing." *Automation in Construction*, 99, 125–139.

Konstantinou, E. and Brilakis, I. (2018). "Matching construction workers across views for automated 3D vision tracking on-site." *Journal of Construction Engineering and Management*, 144(7), 04018061.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks." *Advances in Neural Information Processing Systems*, 25, 1097–1105.

Kropp, C., Koch, C. and König, M. (2018). "Interior construction state recognition with 4D BIM registered image sequences." *Automation in Construction*, 86, 11–32.

Kumar, S. S., Abraham, D. M., Jahanshahi, M. R., Iseley, T. and Starr, J. (2018). "Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks." *Automation in Construction*, 91, 273–283.

Lattanzi, D. and Miller, G. (2017). "Review of robotic infrastructure inspection systems." *Journal of Infrastructure Systems*, American Society of Civil Engineers, 23(3), 04017004.

Lee, Y.-J. and Park, M.-W. (2019). "3D tracking of multiple onsite workers based on stereo vision." *Automation in Construction*, 98, 146–159.

Li, S., Yuan, C., Liu, D. and Cai, H. (2016). "Integrated processing of image and GPR data for automated pothole detection." *Journal of Computing in Civil Engineering*, 30(6), 04016015.

Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., Yu, G., Lu, H., Wei, Y. and Sun, J. (2019). "Rethinking on multi-stage networks for human pose estimation." *arXiv:1901.00148 [cs]*.

Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y. and Sun, J. (2018). "DetNet: A backbone network for object detection." *arXiv:1804.06215 [cs]*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C. (2016). "SSD: Single shot multiBox detector." *arXiv:1512.02325 [cs]*, 9905, 21–37.

Lowe, D. G. (2004). "Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision*, 60(2), 91–110.

Luo, H., Liu, J., Fang, W., Love, P. E. D., Yu, Q. and Lu, Z. (2020). "Real-time smart video surveillance to manage safety: A case study of a transport mega-project." *Advanced Engineering Informatics*, 45, 101100.

Luo, H., Xiong, C., Fang, W., Love, P. E. D., Zhang, B. and Ouyang, X. (2018). "Convolutional neural networks: Computer vision-based workforce activity assessment in construction." *Automation in Construction*, 94, 282–289.

Luo, H., Wang, M., Wong, P. K.-Y. and Cheng, J. C. P. (2020). "Full body pose estimation of construction equipment using computer vision and deep learning techniques." *Automation in Construction*, 110, 103016.

Luo, X., Li, H., Cao, D., Dai, F., Seo, J. and Lee, S. (2018). "Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks." *Journal of Computing in Civil Engineering*, 32(3), 04018012.

Mahony, N. O., Campbell, S., Carvalho, A., Harapanahalli, S., Velasco-Hernandez, G., Krpalkova, L., Riordan, D. and Walsh, J. (2020). "Deep learning vs. traditional computer vision." *arXiv:1910.13796 [cs]*, 943.

Marcus, G. (2018). "Deep learning: A critical appraisal." *arXiv:1801.00631 [cs, stat]*.

McLaughlin, E., Charron, N. and Narasimhan, S. (2020). "Automated defect quantification in concrete bridges using robotics and deep learning." *Journal of Computing in Civil Engineering*, 34(5), 04020029.

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." *Automation in Construction*, 32, 24–37.

Menendez, E., Victores, J. G., Montero, R., Martínez, S. and Balaguer, C. (2018). "Tunnel structural inspection and assessment using an autonomous robotic system." *Automation in Construction*, 87, 117–126.

Merrill, M. (2021). ForConstructionPros, < https://www.forconstructionpros.com/construction-technology/article/21451673/construction-digitization-understanding-the-impact-on-profitability > (accessed 10 November 2021).

Mneymneh, B. E., Abbas, M. and Khoury, H. (2019). "Vision-based framework for intelligent monitoring of hardhat wearing on construction sites." *Journal of Computing in Civil Engineering*, 33(2), 04018066.

Morgenthal, G., Hallermann, N., Kersten, J., Taraben, J., Debus, P., Helmrich, M. and Rodehorst, V. (2019). "Framework for automated UAS-based structural condition assessment of bridges." *Automation in Construction*, 97, 77–95.

Nanni, L., Ghidoni, S. and Brahnam, S. (2017). "Handcrafted vs. non-handcrafted features for computer vision classification." *Pattern Recognition*, 71, 158–172.

Oliveira, H. and Correia, P. L. (2014). "CrackIT — An image processing toolbox for crack detection and characterization." *2014 IEEE International Conference on Image Processing (ICIP)*, 798–802.

Ouma, Y. O. and Hahn, M. (2017). "Pothole detection on asphalt pavements from 2D-colour pothole images using fuzzy c-means clustering and morphological reconstruction." *Automation in Construction*, 83, 196–211.

Park, M.-W., Koch, C. and Brilakis, I. (2012). "Three-dimensional tracking of construction resources using an onsite camera system." *Journal of Computing in Civil Engineering*, 26(4), 541–549.

Park, M.-W. and Brilakis, I. (2016). "Continuous localization of construction workers via integration of detection and tracking." *Automation in Construction*, 72, 129–142.

Pan, Y., Zhang, G. and Zhang, L. (2020). "A spatial-channel hierarchical deep learning network for pixel-level automated crack detection." *Automation in Construction*, 119, 103357.

Pour Rahimian, F., S. Seyedzadeh, S. Oliver, S. Rodriguez, and N. Dawood. 2020. "On-demand monitoring of construction projects through a game-like hybrid application of BIM and machine learning." *Automation in Construction*, 110: 103012. https://doi.org/10.1016/j.autcon.2019.103012.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016). "You only look once: Unified, real-time object detection." *arXiv:1506.02640 [cs]*.

Redmon, J. and Farhadi, A. (2016). "YOLO9000: Better, faster, stronger." *arXiv:1612.08242 [cs]*.

Redmon, J. and Farhadi, A. (2018). "YOLOv3: An incremental improvement." *arXiv:1804.02767 [cs]*.

Ren, S., He, K., Girshick, R. and Sun, J. (2016). "Faster R-CNN: Towards real-time object detection with region proposal networks." *arXiv:1506.01497 [cs]*.

Rezazadeh Azar, E. and McCabe, B. (2012). "Automated visual recognition of dump trucks in construction videos." *Journal of Computing in Civil Engineering*, 26(6), 769–781.

Rezazadeh Azar, E. and McCabe, B. (2012b). "Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos." *Automation in Construction*, 24, 194–202.

Rezazadeh Azar, E. (2016). "Construction equipment identification using marker-based recognition and an active zoom camera." *Journal of Computing in Civil Engineering*, 30(3), 04015033.

Roberts, D., Torres, C. W., Tang, S. and Golparvar-Fard, M. (2020). "Vision-based construction worker activity analysis informed by body posture." *Journal of Computing in Civil Engineering*, 34(4), 04020017.

Rodriguez-Gonzalvez, P., Gonzalez-Aguilera, D., Lopez-Jimenez, G. and Picon-Cabrera, I. (2014). "Image-based modeling of built environment from an unmanned aerial system." *Automation in Construction*, 48, 44–52.

Rosnell, T. and Honkavaara, E. (2012). "Point cloud generation from aerial image data acquired by a quadrocopter type micro unmanned aerial vehicle and a digital still camera." *Sensors*, Molecular Diversity Preservation International, 12(1), 453–480.

Shen, H.-K., Chen, P.-H. and Chang, L.-M. (2018). "Human-visual-perception-like intensity recognition for color rust images based on artificial neural network." *Automation in Construction*, 90, 178–187.

Siebert, S. and Teizer, J. (2014). "Mobile 3D mapping for surveying earthwork projects using an Unmanned Aerial Vehicle (UAV) system." *Automation in Construction*, 41, 1–14.

Simonyan, K. and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition." *arXiv:1409.1556 [cs]*.

Soltani, M. M., Zhu, Z. and Hammad, A. (2017). "Skeleton estimation of excavator by detecting its parts." *Automation in Construction*, 82, 1–15.

Soltani, M. M., Zhu, Z. and Hammad, A. (2018). "Framework for location data fusion and pose estimation of excavators using stereo vision." *Journal of Computing in Civil Engineering*, 32(6), 04018045.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2014). "Going deeper with convolutions." *arXiv:1409.4842 [cs]*.

Tang, S., Golparvar-Fard, M., Naphade, M. and Gopalakrishna, M. M. (2020). "Video-based motion trajectory forecasting method for proactive construction safety monitoring systems." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 34(6), 04020041.

Tang, S., Roberts, D. and Golparvar-Fard, M. (2020). "Human-object interaction recognition for automatic construction site safety inspection." *Automation in Construction*, 120, 103356.

Ullman, S. (1979). "The interpretation of structure from motion." *Proceedings of the Royal Society of London. Series B. Biological Sciences*, Royal Society, 203(1153), 405–426.

Wang, L., Kawaguchi, K. and Wang, P. (2020). "Damaged ceiling detection and localization in large-span structures using convolutional neural networks." *Automation in Construction*, 116, 103230.

Wang, Z., Li, H. and Zhang, X. (2019). "Construction waste recycling robot for nails and screws: Computer vision technology and neural network approach." *Automation in Construction*, 97, 220–228.

Wei, S.-E., Ramakrishna, V., Kanade, T. and Sheikh, Y. (2016). "Convolutional pose machines." *arXiv:1602.00134 [cs]*.

Wiley, V. and Lucas, T. (2018). "Computer vision and image processing: A paper review." *International Journal of Artificial Intelligence Research*, 2(1), 29–36.

Xie, R., Yao, J., Liu, K., Lu, X., Liu, Y., Xia, M. and Zeng, Q. (2018). "Automatic multi-image stitching for concrete bridge inspection by combining point and line features." *Automation in Construction*, 90, 265–280.

Xu, J. and Yoon, H.-S. (2019). "Vision-based estimation of excavator manipulator pose for automated grading control." *Automation in Construction*, 98, 122–131.

Yoder, L. and Scherer, S. (2016). "Autonomous exploration for infrastructure modeling with a micro aerial vehicle." *Field and Service Robotics: Results of the 10th International Conference*, Springer Tracts in Advanced Robotics, D. S. Wettergreen and T. D. Barfoot, eds., Springer International Publishing, Cham, 427–440.

Yan, X., Zhang, H. and Li, H. (2020). "Computer vision-based recognition of 3D relationship between construction entities for monitoring struck-by accidents." *Computer-Aided Civil and Infrastructure Engineering*, 35(9), 1023–1038.

Yuan, C., Li, S. and Cai, H. (2017). "Vision-based excavator detection and tracking using hybrid kinematic shapes and key nodes." *Journal of Computing in Civil Engineering*, 31(1), 04016038.

Zhang, H., Yan, X. and Li, H. (2018). "Ergonomic posture recognition using 3D view-invariant features from single ordinary camera." *Automation in Construction*, 94, 1–10.

Zhang, M., Zhu, M. and Zhao, X. (2020). "Recognition of high-risk scenarios in building construction based on image semantics." *Journal of Computing in Civil Engineering*, 34(4), 04020019.

Zhu, Z., Ren, X. and Chen, Z. (2016). "Visual tracking of construction jobsite workforce and equipment with particle filtering." *Journal of Computing in Civil Engineering*, 30(6), 04016023.

Zuo, X., Dai, B., Shan, Y., Shen, J., Hu, C. and Huang, S. (2020). "Classifying cracks at sub-class level in closed circuit television sewer inspection videos." *Automation in Construction*, 118, 103289.