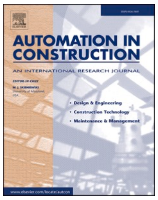
Automation in Construction 129 (2021) 103817

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon



Vision-based method of automatically detecting construction video highlights by integrating machine tracking and CNN feature extraction

Bo Xiao [a], Xianfei Yin [b], Shih-Chung Kang [a,*]

[a] Department of Civil and Environmental Engineering, University of Alberta, Edmonton T6G 1H9, Canada
[b] Department of Construction Management and Engineering, University of Twente, Enschede, the Netherlands.

ARTICLE INFO

Keywords:
Video highlight detection
Keyframe extraction
Object tracking
Deep learning
Productivity analysis

ABSTRACT

Automatic analysis of construction video footage is beneficial for project management tasks such as productivity analysis and safety control. However, construction videos are usually long in duration and only contain limited useful information to engineers, while the storage of video data from construction projects is challenging. To obtain and store useful video footage systematically and concisely, this research proposes a vision-based method to automatically generate video highlights from construction videos. The proposed approach is validated through two case studies: a gate scenario and an earthmoving scenario. In experiments, the proposed method has achieved 89.2% on precision and 93.3% on recall, which outperforms the feature-based method by 12.7% and 17.2% on precision and recall, respectively. Meanwhile, the proposed method reduces the required digital storage space by 94.6%. The proposed approach offers potential benefits to construction management in terms of significantly reducing video storage space and efficiently indexing construction video footage.

1. Introduction

Cameras have emerged as an important piece of equipment in construction management, widely used for remote monitoring of job sites. In surveying 142 construction experts, Bohn and Teizer [1] identified that construction cameras can efficiently reduce project budgets in terms of communication, resource management, and site security. Indeed, construction videos contain important visual information that can serve multiple purposes in project management, such as crew productivity evaluation [2], material logistics management [3], and safety control [4]. As such, systematic storage of construction video footage is critical with respect to the retrieval, analysis, and documentation of construction activities throughout the project life cycle.

Despite offering a range of potential benefits, the use of raw construction video footage is challenged in two notable respects. First, retrieval of the desired information from unstructured construction video footage is time-consuming and labor-intensive because construction videos have a long duration and only a few clips contain useful project information [5]. In this regard, some owners and engineers underestimate the value and utility of construction videos due to the difficulty of browsing these videos. Second, the sheer volume of video footage generated from continuously recording construction sites can become unmanageable. For instance, a one-hour video in 1080p resolution necessitates approximately 2 GB of digital storage space. Assuming one camera streams 3000 h in a one-year construction project, 6000 GB of space is required for storing this footage. As such, many project managers prefer to delete videos at one- or two-week intervals, resulting in a loss of video records that may be of some value for future reference.

Construction videos contain a significant number of redundant frames that can be potentially eliminated without losing the relevant project management information [6]. For example, the footage captured during non-working hours has negligible value for management purposes. Even in working hours, most video clips are useless when project progress is slow. By removing these unnecessary frames, processed video can replace the raw construction videos for productivity analysis, logistics management, and safety control. By attaching the time stamp and content information (e.g., objects and activities), the condensed video can be stored economically, more easily indexed, and efficiently retrieved for project management purposes.

Video highlight detection is a technology refers to the process of creating a summary of important video clips from the original video, and the generated video highlights should have three properties including minimum repetition, representativeness, and diversity [7]. The video

* Corresponding author.
E-mail addresses: bxiao2@ualberta.ca (B. Xiao), x.yin@utwente.nl (X. Yin), sckang@ualberta.ca (S.-C. Kang).

https://doi.org/10.1016/j.autcon.2021.103817
Received 7 January 2021; Received in revised form 13 June 2021; Accepted 15 June 2021
Available online 20 June 2021
0926-5805/© 2021 Elsevier B.V. All rights reserved.

highlights allow users to obtain certain perspectives of a video without having to view the raw footage in its entirety. This technology has enjoyed success in the entertainment field (e.g., sports highlights and films) [8]. In construction, video highlight detection can be used to "distill" the raw construction videos and help project managers to quickly understand the salient developments at a given job site. Generally, highlight detection methods select keyframes based on image feature changes and then combine clips around keyframes to produce video highlights. However, for three reasons in particular, these feature-based methods are not able to efficiently detect construction video highlights: (1) construction videos have frequent illumination changes, which decrease the highlights detection performance; (2) keyframes in construction cannot be simply defined as the frames with image features change rapidly; and (3) video highlights are expected to be interpretable and flexible for construction management.

To address these issues, this paper proposes a vision-based method to detect video highlights from construction videos. The proposed method explores the context information from videos by tracking construction machines, and then selects object keyframes by analyzing the content information as prescribed by pre-defined construction rules. In parallel, convolutional neural networks (CNN) are employed to extract features from each frame, while the feature keyframes can be selected by calculating the feature changes. As such, the object keyframes and feature keyframes can be processed to produce video highlights. By combining the context and feature information, the proposed method can generate accurate and representative video highlights to replace the original raw construction video footage. This research is expected to help project managers to efficiently retrieve and economically store their job site video footage.

2. Literature review

This section presents a review of the literature related to the detection of video highlights in construction video footage. First, work related to the use of video recording to monitor construction sites is introduced. Second, the state-of-the-art of video highlight detection methods is reviewed, addressing both the mechanism and the various applications that have been reported. Finally, a comprehensive review of the progress made with respect to applying video highlight detection in construction engineering is presented.

2.1. Video monitoring in construction

Video monitoring of sites has become increasingly popular in construction management in recent years, as it allows project managers to monitor the status of their job sites remotely [9]. Compared with other monitoring technologies (e.g., GPS, radio frequency identification, laser scanner), the use of cameras offers the advantages of lower cost, simple installation and maintenance, and larger monitoring range [10]. Cameras can also reduce project budgets in terms of communication, resource planning, and site security [1]. In these respects, cameras are versatile tools in construction engineering for delivering high-quality and more economical projects.

Construction video footage contains important visual information that can be used for productivity analysis, safety management, and carbon footprint monitoring [11]. For instance, Roberts and Golparvar-Fard [12] proposed an end-to-end framework to calculate excavator productivity based on video footage of earthmoving operations. To facilitate site safety, Chi and Caldas [13] tracked construction machines from videos to prevent potential collisions. To minimize environmental impact during construction, Heydarian et al. [14] benchmarked the carbon footprint of construction machines in earthmoving projects using a vision-based method. Besides the abovementioned benefits, construction videos are naturally easily understood and interpreted by humans, and are widely adopted as a form of official project documentation [15]. However, with the increasing use of video monitoring in construction, the efficient storage and use of the large volumes of video footage that result from video monitoring of sites has emerged as an important challenge in construction research.

2.2. Video highlight detection methods

Video highlight detection methods have been researched and used in various applications, including sports highlights, film industry, and video surveillance [16]. For example, Merler et al. [17] developed multimodal excitement features to generate video highlights from a golf tournament (2017 Masters) and two international tennis tournaments (2017 Wimbledon and U.S. Open), where the results having been closely aligned with the official video highlights. Wang et al. [18] proposed a contrastive attention module as the feature representations to produce trailers from full-length movies. Kumar and Shrimankar [19] have proposed a novel deep learning method to extract video highlights from multi-view surveillance videos. Furthermore, Kumar et al. [20] have introduced the self-organizing map into the video highlights detection in video surveillance.

A typical video highlight detection method extracts features from raw videos and then selects keyframes by analyzing changes in the feature space across frames. The video clips around keyframes, usually several seconds, are combined to produce the video highlights for users [8]. Feature extraction and keyframe selection are the main focuses in the computer vision community. A large number of features have been studied for the task of video highlight detection. For example, Laganière et al. [21] integrated the spatio-temporal Hessian matrix to collect image features for video highlight detection. Liu et al. [22] adopted the scale-invariant feature transform (SIFT) to identify the boundary of video highlights. The deep neural network has also emerged as a promising method for extracting features from images by learning from human-created datasets. Mahasseni et al. [23] employed the long short-term memory network (LSTM) to summarize video highlights. Hussain et al. [24] have integrated the CNN and bi-directional LSTM for video summarization using cloud platform for computationally intensive processing. Following this, in an advanced research, Hussain et al. [25] implemented a CNN based video highlights detection method on the Internet-of-things platform. Kumar and Shrimankar [26] have adopted CNN technology to detect video highlights from multi-view videos. Kumar [27] has integrated CNN and optimal local alignment strategy for video highlights detection on the cloud.

Keyframes are a set of representative frames in videos that define the quality of the video highlights. One approach in this regard has been to calculate the Euclidean distance of every two continuous frames. The keyframes can then be identified as the points in the video footage where feature distance changes rapidly [28]. Furthermore, researchers have investigated the feasibility of using machine learning techniques (e.g. clustering, boosting, and graph networks) for keyframe selection. For instance, Kumar and Shrimankar [19] have integrated the AdaBoost [29] for keyframe selection for surveillance videos Mundur et al. [30] developed a keyframe selection method based on Delaunay clustering. Kumar [31] has proposed a keyframe selection method based on the similarity graph. Kumar and Shrimankar [32] have integrated the scale-free network for keyframe selection in video highlights detection. Other studies have employed a method of selecting keyframes by ranking all frames with a pre-defined importance score, such as entropy [33], context prediction score [8], or influence metric [34]. Table 1 summarizes the information of the existing state-of-the-art methods in terms of the published year, methodology mechanism, and evaluation results.

2.3. Video highlight detection in construction

Video highlight detection is an effective solution for reducing redundant video footage [6] and manual inspection effort [44] in construction management. By filtering raw videos, the generated video highlights usually occupy around 10% of the original storage space,

Table 1
Summary of existing video highlights detection in computer vision.

Highlights detection study	Published year	If supervised learning	Mechanism of methodology	Evaluation results		
				Metric	Testing dataset	Results
Zhang et al. [35]	2016	Yes	Integrating LSTM + determinantal point process	F1 score	SumMe [36]	4.18%
					TVSum [37]	58.7%
Mahasseni et al. [23]	2017	No	Integrating GAN + summary-length, diversity, and keyframe regularization	F1 score	SumMe	39.1%
					TVSum	51.5%
					Open Video [38]	72.8%
					VSUMM [39]	60.1%
Kumar and Shrimankar [19]	2018	No	Integrating CNN + AdaBoost	F1 score	Lobby [40]	88.8%
					Office [40]	86.7%
					BL-7F [40]	86.4%
Muhammad et al. [33]	2018	No	Integrating CNN + memorability prediction + entropy score calculation	F1 score	VSUMM	76.0%
Jiao et al. [41]	2018	Yes	Integrating CNN+ 3D attention module+ ranking module	mAP	YouTube Highlights [42]	68%
					SumMe	62%
Kumar [31]	2019	No	Integrating CNN + similarity graph+ highly connected sub-graphs	F1 score	Open Video	64.9%
					VSUMM	52.4%
Kumar and Shrimankar [32]	2019	No	CNN + scale-free network	F1 score	Open Video	63.3%
					VSUMM	51.5%
Xiong et al. [43]	2019	Yes	CNN + paired-wise loss	mAP	YouTube Highlights	56.4%
					TVSum	56.3%
Wang et al. [18]	2020	No	3D CNN + CO-attention	mAP	YouTube Highlights	69.1%
					TVSum	62.8%
Kumar [27]	2021	No	CNN + local alignment strategy	F1 score	Lobby	89.4%
					Office	91.9%
					BL-7F	88.7%

making it more feasible to document the entire construction lifespan without an unreasonable digital storage burden [45]. In current practice, engineers must engage in the time-consuming process of manually browsing entire construction videos in order to observe what is happening on a job site in a specific time period. Moreover, the common practice of having multiple cameras monitoring one construction site adds to the workload of manual inspection [46]. By using video highlights, project managers can conveniently browse major activities occurring on the job site.

The video highlights detection in construction scenarios is different from general scenarios (e.g., entertainment videos, sport videos, and films) in terms of: (1) the visual characteristics of construction videos are special. Construction videos have more unexpected illumination changes than general videos (e.g., entertainment videos) because most construction projects are in an outdoor environment; (2) the definition of keyframes in construction videos is special. The keyframes in construction videos should be selected according to the needs of project management instead of simply defined as frames where visual features change dramatically; and (3) the usages of construction video highlights are special. The video highlights in construction are used for documentation and inspection purposes instead of simply for watching. Therefore, video highlights in construction are expected to be searchable, manageable, and filterable.

Researchers have put a significant amount of effort into developing video highlight detection methods to accommodate construction video characteristics. For instance, Chen and Wang [47] developed construction-specific color, texture, and gradient features for extracting keyframes from videos. The developed methods were tested on four construction videos, and the experimental results suggested that color features generally outperform gradient and texture features. However, that study focused on exploring image features and did not utilize the content information of construction videos. Ham and Kamari [48] proposed a content-based keyframe selection method for construction videos captured by drones that scores frames individually based on the spatial composition of the identified objects. However, their method was designed for drone videos, whereas it cannot be directly applied to videos captured by fixed-position cameras.

2.4. Research gaps and objectives

The related works indicate that video highlights detection methods have achieved huge success in the computer vision community. Most existing methods extract features from videos and then select keyframes by evaluating feature changes across frames. The gaps of existing methods and the objectives of this research have been presented in this section.

2.4.1. Research gaps

Three research gaps have been identified in applying existing feature-based methods to detect video highlights in construction scenarios.

- The performance of existing methods needs to be improved. Unexpected illumination changes in construction videos decrease the performance of feature-based methods. Meanwhile, the feature-based methods define the keyframes as frames where visual feature changes dramatically, which is not related to construction management. As consequence, the detected highlights produced by existing methods are less precise.
- Existing methods lack interpretability. For feature-based methods, it is not clear how the video highlights can help to construction management. And video highlights detected by existing methods are not searchable. For example, in a construction project, the managers would like to inspect all video highlights related to excavators, while this need cannot be met because the feature-based methods are lacking interpretability.
- Existing methods lack flexibility. In the same construction scenario, the requirements of video highlights can be different depends on the project stages. For example, for surveillance videos in a construction gate, the video highlights are required to focus on excavators and dump trucks in the earthmoving stage. As the project progressing, the video highlights are required to focus on concrete mixer trucks in the rebar stage. This kind of flexible needs is difficult to be achieved by feature-based methods.

2.4.2. Research objectives

To fill above gaps, this study explores the potential of rule-based methods in video highlights detection in construction scenarios. The research objectives are two-fold:

- Developing a vision-based video highlights detection method that has sufficient performance in construction scenarios. The proposed method explores both context and feature information for selecting keyframes from construction videos. The video highlights produced by the proposed method are expected to be interpretable and flexible to accommodate practical needs in construction management.
- Developing construction rules for keyframe selection. The developed rules should extract keyframes related to site safety, productivity analysis, and logistics management. The construction rules are considered as "blueprint", which can be easily customized, expanded, and implemented by project managers in individual projects.

3. Proposed methodology

A vision-based method for automatically detecting construction video highlights is introduced and described in this section. The proposed method consists of five main modules: machine tracking, rule-based keyframe selection, CNN feature extraction, similarity evaluation, and video editing.

3.1. Overall framework

The overall framework of our approach is depicted in Fig. 1. As shown in the figure, two types of keyframes are involved in generating video highlights: object keyframes and feature keyframes. Object keyframes are the frames that contain important construction management information related to continuous activities (e.g., machines accessing the working zone). Feature keyframes are the frames where the image feature changes significantly because of scene changes (e.g., camera zooming, edited changeover, task changes). In this research, the object keyframes are used to distill the important information from video clips in which construction machines appeared, while the feature keyframes are used to identify notable developments on the site by scanning the entire video.

First, the input video is processed by the machine tracking module to produce the tracking results, including machine categories, machine identification (ID), and the corresponding pixel locations of machines at each individual frame. A multiple object tracking method, called construction machine tracker (CMT), is adopted for the tracking module. The tracking results are stored in a database, and can be conveniently processed by structured query language (SQL). Then, a rule-based method is used to select object keyframes by applying pre-defined construction rules in analyzing the tracking results. These rules are deployed to explore the working zone, working status, and working interaction information of construction machines. For feature keyframe selection, the ResNet50 CNN is employed to extract high-level features from all frames of the input video. The features across frames are evaluated using cosine similarity to select the keyframes that represent scene changes. Finally, object keyframes and feature keyframes are combined together in the video editing module to remove the duplicated keyframes and generate the video highlights. The individual modules are introduced in detail in the following subsections.

3.2. Machine tracking

The machine tracking module tracks construction machines from the input video sequences in order to generate information such as machine categories, IDs, and pixel locations. Tracking construction objects is challenging because of the dynamic environment, occlusions, and illumination changes. A robust tracking method that produces precise bounding boxes of construction machines is the foundation of the object keyframe selection. As mentioned above, CMT [49] is adopted for this

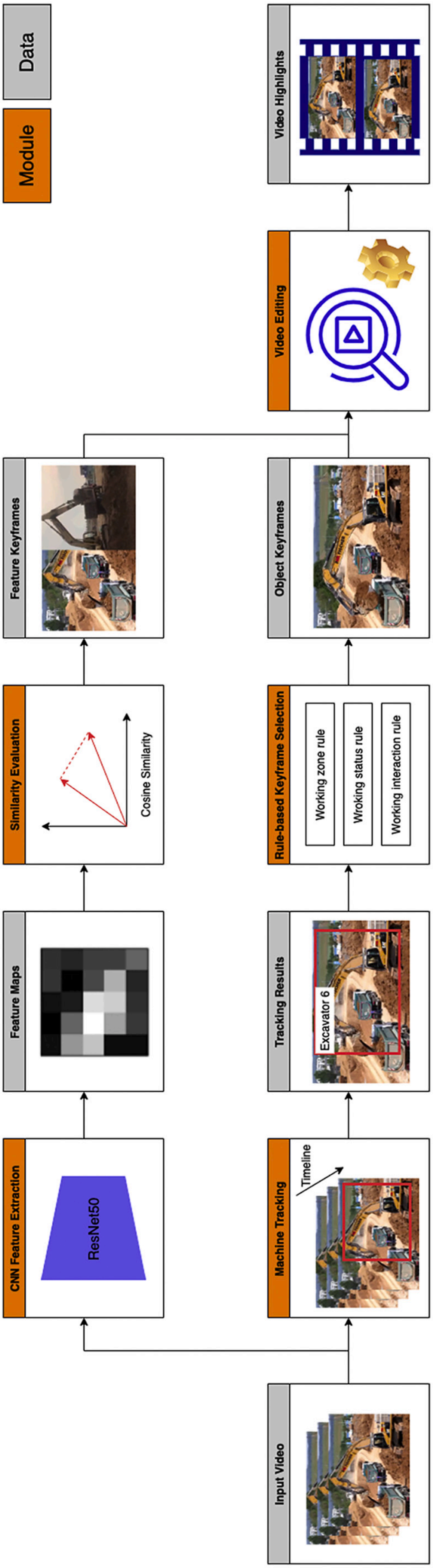


Fig. 1. Overall framework of the proposed method.

purpose because of its robust performance in construction scenarios. CMT was developed specifically for tracking construction machines in complex environments, and it has achieved high tracking robustness (93.2% in multiple-object tracking accuracy) and high processing speed (20.8 frames per second) in experiments.

An overview of the CMT method is provided in Fig. 2. As shown in the figure, images, after being resized to 416 × 416 pixels, are processed by the deep learning detector, YOLO-v3 [50]. Then, the detection results across frames are associated by Intersection over Union (IoU) and image hashing features. According to the association results, the CMT formulates the tracking problem into a linear assignment problem that matches each individual detection bounding box in the current frame with a bounding box in the previous frame. Finally, the linear assignment problem is solved by the Jonker-Volgenant algorithm [51] to produce the tracking results.

The use of an annotated image dataset is crucial for training the YOLO-v3 detector. In the present research, the Alberta Construction Image Dataset (ACID) [52] is adopted. ACID contains ten types of construction machines: excavator, compactor, dozer, grader, dump truck, concrete mixer truck, wheel loader, backhoe loader, tower crane, and mobile crane. It is also important to track commuter cars in some construction scenarios (e.g., construction gate scenario); as such, we randomly select 2000 car images containing 3895 car objects from the COCO dataset to combine with ACID for training purposes.

A database is created to store the tracking results. The table contains nine attributes: frame number, time stamp, if_tracked, machine category, machine ID, *cx*, *cy*, *w*, and *h*. The frame number attribute indicates the sequencing of the current frame, and the time stamp attribute shows the time of the current frame in the video, accurate to the second. Meanwhile, the if_tracked attribute is a Boolean value that indicates whether any machine has been identified in the frame. If there is a machine object in the current frame, the type of machine and its ID number will be stored in the machine category attribute and the machine ID attribute, respectively. The *cx* and *cy* attributes indicate the pixel coordinates of the centroid point of the bounding box, while the *w* and *h* attributes refer to the width and height of the machine bounding box, respectively. By using database, the tracking results can be organized in a structured format within the database and conveniently analyzed by the rule-based keyframe detection module.

3.3. Rule-based Keyframe detection

The purpose of this module is to select object keyframes by integrating predefined construction rules and tracking results. Three types of construction rules are proposed—working zone rule, working status rule, and working interaction rule—where Table 2 summarizes the definition and target of each rule.

3.4. Working zone rule

Working zone control is important for site safety and resource logistics in construction management. For instance, there is a risk of collisions between machines and pedestrians when machines access the working zone in some scenarios (e.g., road maintenance construction). The time stamp of machines accessing the working zone also indicates the actual scheduling information that can be compared with the planned schedules for logistics management purposes. Therefore, the frames that feature interested machines entering or leaving the working zone are selected as keyframes in the present study.

Table 2
Summary of predefined construction rules.

Rule name	Rule definition	Rule target
Working zone rule	Any frame that contains interested machines entering or leaving the working zone should be considered a keyframe.	Site safety and logistics management
Working status rule	Any frame that contains machine working status changes between working and idling in the working zone should be considered a keyframe.	Productivity analysis
Working interaction rule	Any frame that contains extensive overlap between cooperating machines in the working zone should be considered a keyframe.	Site safety and productivity analysis

Eq. 1 shows the judgement criterion underlying the working zone rule, where A_{ABCD} is the area of the working zone polygon ABCD, P_i is the pixel location of the machine object's central point in frame *i*, and *fr* is the frame rate of the video. Connecting the location of the central point at the current frame, *i*, and the location at frame $i - fr$ can generate a segment P_iP_{i-fr}. If the segment P_iP_{i-fr} has more than 0 intersections with the polygon area A_{ABCD}, frame *i* is selected as the keyframe. Fig. 3 shows an example of application of the working zone rule. In Fig. 3(a), segment P_iP_{i-fr} has no intersection with the working zone ABCD and should be ignored for keyframe selection. In Fig. 3(b), the dump truck is entering the working zone, while segment P_iP_{i-fr} has one intersection with the working zone ABCD. Therefore, this frame should be selected as a keyframe based on the working zone rule.

$$\mathrm{Count}\left(P_iP_{i-fr} \cap A_{ABCD}\right) > 0 \tag{1}$$

3.5. Working status rule

Identification of the frames that contain working status changes of construction machines is an essential task for productivity analysis, as this information can be used to automatically calculate the machine idling time and efficiency factor. The working status rule selects keyframes in which the status of the interested machine changes from idling to working or from working to idling. This rule is only interested in the machine status changes occurring in the working zone (i.e., the centroid point of the machine object must be in the working zone). When a machine is idling, it should be noted, the pixel location of this object may change slightly because of the tracking bounding box precision.

The judgement criterion underlying the working status rule at frame *i* is defined as per Eq. 2, where cx_i and cy_i represent the *x* and *y* coordinates of the central point of the machine object, respectively, *fr* is the video frame rate, and $k \in \mathrm{N}$. Eq. 2 calculates the average distance between the central points in the current frame, *i*, and the previous frame, $i - 1$, in *fr* continuous frames. When the average distance is greater than d_1, the machine status is considered to be "working". The machine status is "idling", meanwhile, if the average distance is less than d_2. When the average distance is between d_1 and d_2, the current frame indicates the machine is in transition between working and idling status, and as such it should be selected as a keyframe. The variables d_1 and d_2 are threshold values and need to be set for the given construction scenario.

$$d_1 > \frac{1}{fr}\sum_{k\in(i-fr,fr]} \sqrt[2]{\left(cx_k - cx_{k-1}\right)^2 + \left(cy_k - cy_{k-1}\right)^2} > d_2 \tag{2}$$

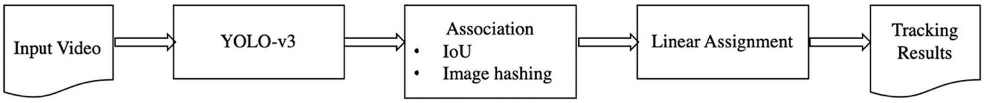
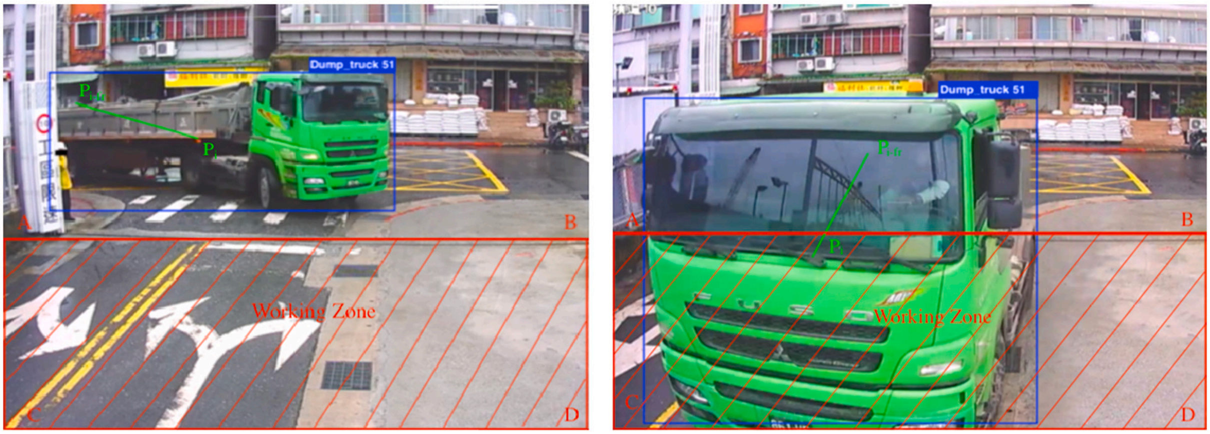


Fig. 2. Overview of CMT method.



(a) ignored frame (no intersection)

(b) keyframe (intersection with working zone)

Fig. 3. Example of keyframe selection applying the working zone rule.

3.6. *Working interaction rule*

A high level of interaction between two construction machines is often indicative a meaningful moment with respect to crew productivity analysis and safety monitoring. For example, high overlap between the excavator and the dump truck in earthmoving represents a loading activity, which can be used for cyclic productivity calculation. High overlap between two dump trucks, meanwhile, may signify a potential collision and may be of interest for safety alerting purposes. In our research, the working interaction rule selects keyframes by analyzing the overlap between two interested construction machines in the working zone. To apply the working interaction rule, the IoU between two machine objects, m and n, at the frame i is calculated by means of Eq. 3, as illustrated in Fig. 4. If the average IoU in fr continuous frames (see Eq. 4) is greater than threshold a ($k \in$ N), the current frame is considered a keyframe. Fig. 5 shows an example of an application of the working interaction rule. In Fig. 5(a), the excavator and the dump truck are overlapping. If these two machines are in the working zone and the average IoU is greater than a, this frame should be selected as the object keyframe. In Fig. 5(b), the excavator and the dump truck have no interactions, so this frame will not be selected as a keyframe.

$$IoU(m,n) = \frac{Area(m) \cap Area(n)}{Area(m) \cup Area(n)} \tag{3}$$

$$\frac{1}{fr} \sum_{k \in (i-fr,i]} IoU(m_k, n_k) > a \tag{4}$$

To apply the abovementioned rules successfully in construction scenarios, two strategies need to be considered: (1) each type of construction rule should be considered a “blueprint”, where several individual rules can be generated by changing the interested classes of machines (for example, two working interaction rules can be generated in the earthmoving scenario, where one focuses on the excavator and dump truck and another focuses on the wheel loader and dump truck); and (2) it is not necessary to apply all three types of construction rules to the same construction scenario. The procedure for generating individual rules consists of four steps: selecting the type of construction rule, defining the working zone, selecting the interested construction machines, and setting up threshold values (if needed). The keyframes detected by each individual rule are simply combined together and inputted to the video editing module.

3.7. *CNN feature extraction and similarity evaluation*

The CNN feature extraction and similarity evaluation are employed to detect feature keyframes. In the present study, feature keyframes are used for two purposes: (1) to represent video clips that have no machine objects; and (2) as an addition to object keyframes in video clips that do have machine objects, since feature keyframes are more effective than object keyframes for describing scene changes (e.g., camera zooming, moving, and length transition). Compared with manually designed features, such as SIFT, CNN has been shown in previous studies to be more effective in representing construction images [53–55].

In CNN feature extraction, all frames in the construction video are processed with the CNN neural networks to produce feature vectors for the purpose of representing original frames. In this research, the ResNet50 neural network [56] is employed for feature extraction due to its excellent performance in computer vision applications. The ResNet50 has 50 layers of neural networks for implementing the residual block,

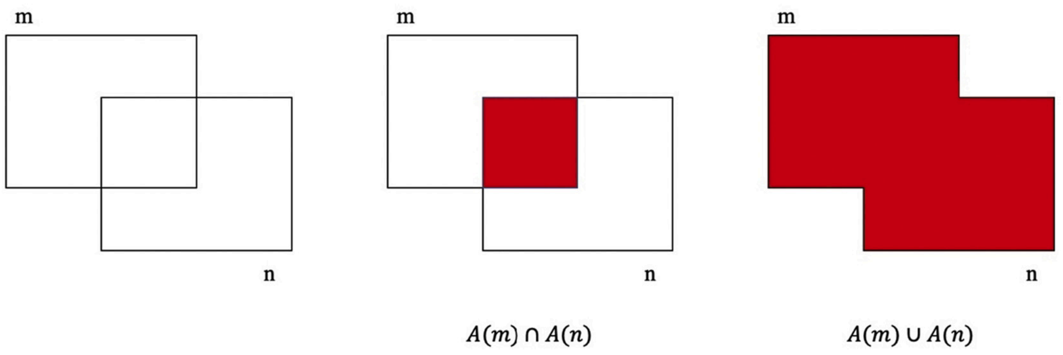
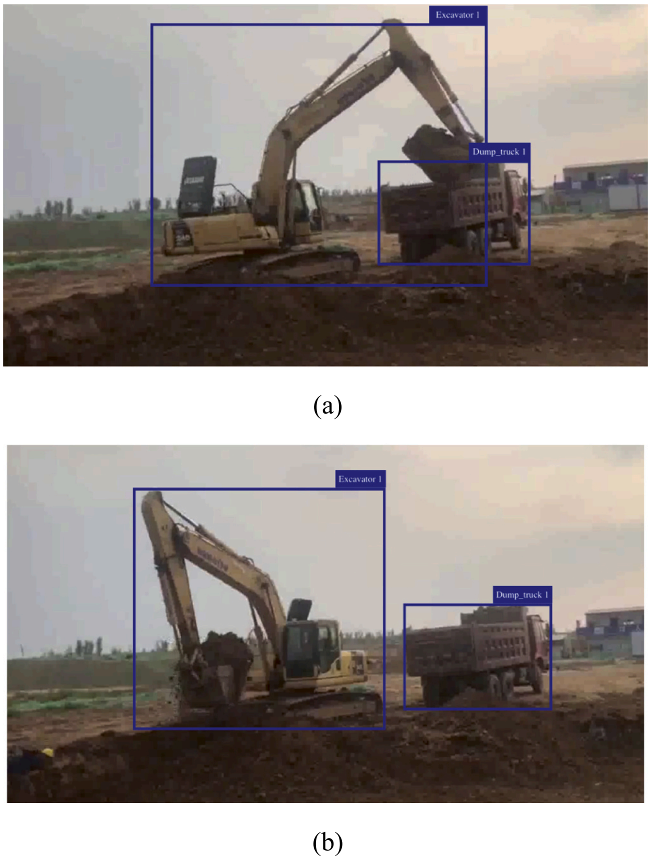


Fig. 4. Illustration of Interaction over Union (IoU).



(a)

(b)

Fig. 5. Example of keyframe selection using working interaction rule.

where the residual block is defined as per Eq. 5.

$$y = \mathscr{F}(\mathrm{X}) + \mathrm{X} \tag{5}$$

where X is the input feature map, $\mathscr{F}(\mathrm{X})$ is the feature map processed by the stacked layers, and y is the output feature map of the residual block.

As shown in Fig. 6, the residual block is a "shortcut connection" that adds the outputs of the stacked layer $\mathscr{F}(\mathrm{X})$ to the input feature map X, where this residual learning solves the gradient vanishing problem in training the deep neural networks. In the CNN feature extraction module, all frames of the input video are first resized into 224 × 224 resolution. The resized frames are then inputted to the ResNet50, which has been pretrained on the ImageNet dataset [57] for forward propagation. A vector with dimensions of 2048 × 1 can then be extracted from the flatten layer as the output of this module.

The purpose of the similarity evaluation module is to select feature keyframes based on the average feature similarity $AS(i)$ at each frame. To calculate AS, we firstly define the similarity $S(a,b)$ of two frames (i.e. a and b) as the cosine similarity [58] of their corresponding feature vectors (as shown in Eq.6).

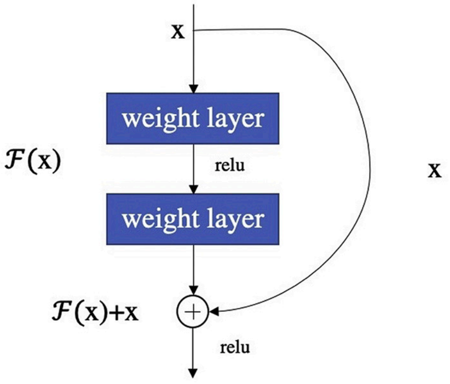


Fig. 6. Illustration of residual block.

$$S(a,b) = \frac{v(a)v(b)^{'}}{\|v(a)\|\|v(b)\|} \tag{6}$$

where $v(a)$ and $v(b)$ are the feature vectors processed by ResNet50 for frame a and frame b, respectively, and $\|v(a)\|$ is the norm of vector $v(a)$.

Then, the average feature similarity $AS(i)$ at frame i (defined in Eq.7) is calculated as the average similarity between the feature vectors of the current frame i and the frame $(i - fr)$ in one continuous second where $k \in \mathrm{N}$.

$$AS(i) = \frac{1}{fr}\sum_{k\in(i-fr,i]} S(k, k - fr) \tag{7}$$

If $AS(i)$ is smaller than threshold value s, the current frame i is considered to be a feature keyframe. Here, the smaller the value of s that is adopted, the fewer feature keyframes will be detected. In construction videos, continuous frames usually have high similarity because construction activities change in a relatively gradual manner. In the present case, the threshold s, at just 0.9, is relatively small. Because the role of the similarity evaluation module is to detect significant feature changes resulting from scene changes.

3.8. Video editing

The function of the video editing module is to produce video highlights based on detected object keyframes and video keyframes. This is carried out in two steps: redundancy removal and video concatenation. It should be noted that the detected object keyframes and feature keyframes are intervals of sets of frames rather than discrete frames. The object keyframes can be represented as $T_{object} = \{[s,e]_1, [s,e]_2, \dots, [s,e]_i\}$, where $[s,e]_i$ is a time interval of keyframes, and s and e are the start- and end-frame number of the time interval, respectively. It is possible that the time interval may have only a few frames due to tracking errors. As such, any time intervals that have fewer than five frames $(e - s < 5)$ are first removed. To generate useful and understandable video highlights, each video clip should be several seconds in length at a minimum in order for users to understand what is occurring in the highlight. In consideration of this, we expand the time interval $[s,e]_i$ to $[s',e']_i$ as per Eq. 8. This equation calculates the median frame of the time interval $[s, e]_i$ and then finds the n seconds before and after the central frame as the basis for determining the new time interval, where the present research assigns n a value of 2. After this step, all time intervals have the same length of $4fr$. It is possible that the different construction rules will locate adjacent, overlapping, or identical keyframes. In other words, many time intervals in T_{object} are redundant and will need to be removed. For two continuous time intervals, we remove the first interval $[s',e']_i$ if $s_{i+1}' - s_i' \leq fr$. If two continuous time intervals are close to one another $(2n \times fr > s_{i+1}' - s_i' > fr)$, they are merged to a new interval $[s_i' - n \times fr, e_{i+1}' + n \times fr]$. The same process is conducted with respect to the feature keyframes $T_{feature}$.

$$[s^{'}, e^{'}]_i = \left[floor\left(\frac{s+e}{2}\right) - n \times fr\ , floor\left(\frac{s+e}{2}\right) + n \times fr\right]_i \tag{8}$$

The processed T_{object} and $T_{feature}$ can then be used to produce video highlights by extracting the corresponding frames from the original construction video and concatenated these frames together. It should be noted that the object keyframes and feature keyframes may be overlapping. In the present study, overlapping frames between object keyframes and feature keyframes are not removed. Instead, all object keyframes and feature keyframes are retained in the final video highlights, annotating each keyframe with different colors of symbols. Users are thereby able to recognize whether a given video highlight frame belongs to object or feature keyframes.

4. Implementation and case studies

This section describes the implementation of the proposed method. To validate its feasibility, two case studies have been conducted: a construction gate case and an earthmoving case.

4.1. Implementation

The proposed method is programmed in Python 3.6, and the Opencv library is adopted for the video I/O. In the CMT tracker, the backbone detector YOLO-v3, originally programmed in C, is implemented via Python wrapper with an acceleration of CUDA 9.0 and Cudnn 7.0. Meanwhile, the rule-based keyframe selection module is built using the SQLite-Python library, whereas the construction rules are implemented using SQL queries. The ResNet50 is implemented using the Pytorch library, while the cosine similarity is built using the scikit-kearn library. For video editing, the Moviepy library is employed to generate the final video highlights.The proposed method is tested in an Ubuntu 18.04 64-bit system environment.

For the hardware configuration, the proposed method is tested on a computer with a NVIDIA GTX 1080Ti graphics card, 11 GB memory, an Intel Core i9-7920×@2.90 Hz CPU with 12 cores, and two 32 GB memory cards. The processing speed when implementing the proposed method is approximately 7 frames per second. It should be noted that the graphics card specifications affect the speed of executing YOLO-v3 and ReNet50. As such, the processing speed can be increased by upgrading to an advanced graphics card or implementing parallel programming.

4.2. Case study 1: construction gate

In case 1, the proposed method was tested on construction video footage captured from a gate specifically for machine traffic, with dump trucks, concrete mixer trucks, dozers, and cars all appearing in this footage. The construction gate scenario was adopted as a case study in this research for two reasons: (1) construction gate video footage contains important information about what equipment is present on the construction site at a given time (i.e., timestamped arrivals and departures of construction machines), which is crucial for construction gate control; and (2) the need for video highlights is particularly pressing for gate scenarios since almost all construction gates feature cameras capturing large volumes of raw video footage.

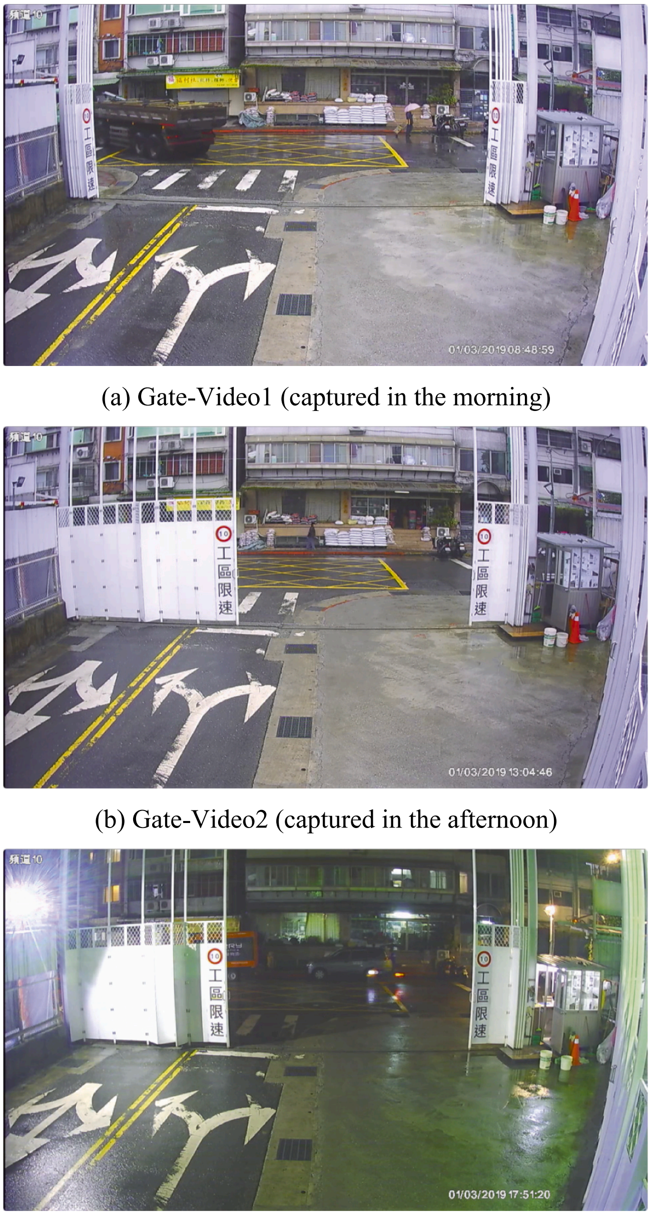
4.2.1. Experimental setup

Three gate videos capturing footage of the same construction gate were used in the experiment. The relevant information regarding the test videos (video duration, resolution, frame rate, and number of highlights) is summarized in Table 3. Fig. 7 shows example images for each test video. Three test videos corresponding to different times of day, i.e., morning, afternoon, and evening, were captured in order to investigate the feasibility of the proposed method under different illumination conditions.

To evaluate the performance of the proposed method, the ground truth video highlights in each test video had to be manually annotated. Of course, the annotation of video highlights is an inherently subjective task since there is no absolute definition of what constitutes a highlight. However, construction engineers are likely to share similar points of view with regard to what constitutes a useful video highlight of construction site footage for construction management purposes based on their experience, knowledge, and intuition. In our research, five graduate students majoring in construction management were invited to manually identify highlights from construction video footage. Their annotations of these highlights consisted of a time stamp of the highlight and a short description (e.g., "From 27:01 to 27:05: A dump truck exits the gate and turns right"). A two-step strategy was implemented for video highlights annotation: (1) each participant was asked to find the video clips in which machines access the construction gate, the camera working state changes, or unusual activities occur, or other clips they think may be highlights. (2) the author of this research manually browses the video highlights annotated by all participants to decide the final video highlights as ground truth. The annotated video highlights were then used to assess the proposed method.

Table 3
Specifications of test videos for construction gate case.

	Duration (minutes)	Video Resolution	Frame rate (fps)	Number of highlights contained
Gate-Video1	60	1920 × 1080	12	20
Gate-Video2	60	1920 × 1080	12	19
Gate-Video3	60	1920 × 1080	12	14



(a) Gate-Video1 (captured in the morning)

(b) Gate-Video2 (captured in the afternoon)

(c) Gate-Video3 (captured in the evening)

Fig. 7. Example images from test videos in construction gate case.

In case 1, the working zone rule and working status rule were applied, with the detailed configurations of these two rules summarized in Table 4. The working zone rule was applied to detect video highlights featuring machines accessing the gate. Machines suddenly stopping in the gate area, meanwhile, signified potential highlights that could be detected by the working status rule.

To validate the feasibility of the proposed method, we removed the machine tracking module and the rule-based keyframe detection module from the proposed method as the baseline. The baseline method retained the same CNN feature extraction module, similarity evaluation module, and video editing module as those in the proposed method. In the baseline method, however, the threshold s in the similarity evaluation module was set at 0.95, a higher value of s than that employed in the proposed method, in order to produce more video highlights. Fig. 8 shows the framework of the baseline method.

4.2.2. Evaluation metrics

Following the protocols set out in previous work [35], precision, recall, and F1 score are employed as the evaluation metrics in the present study, where A denotes the video highlights generated by the proposed method, and B denotes the annotated ground truth video highlights. Precision, meanwhile, is the measurement of how accurate the highlight detection method is (see Eq. 9), while recall measures how effective the highlight detection method is in identifying the correct highlight clips according to Eq. 10. The F1 score is the harmonic mean of precision and recall as defined in Eq. 11.

$$Precision = \frac{Number\ of\ correct\ highlight\ clips}{Number\ of\ highlight\ clips\ in\ A} \tag{9}$$

$$Recall = \frac{Number\ of\ correct\ highlight\ clips}{Number\ of\ highlight\ clips\ in\ B} \tag{10}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

where the correct highlight clip is decided by the temporal intersection of union (TIoU) between the generated highlight clip a ($a \in A$) and the ground truth highlight clip b ($b \in B$), as expressed in Eq. 12. Fig. 9 illustrates the process of computing TIoU. If the value of TIoU is greater than 0.5, the highlight clip a is considered to be a correct highlight clip.

$$TIoU = \frac{duration\ (a) \cap duration(b)}{duration\ (a) \cup duration(b)} \tag{12}$$

4.2.3. Experimental results

Table 5 illustrates the experimental results in terms of precision, recall, F1 score, and the number of correct highlights. The proposed method achieved 87.7% on precision, 94.3% on recall, and 90.8% on F1 score on average, which is 13.8% higher on precision, 14% higher on recall, and 14.1% higher in terms of F1 score compared to the baseline method. Meanwhile, the proposed method detected 1.4 more correct video highlights from each video compared to the baseline method. The experimental results indicate that the proposed method is more robust and precise than the feature-based highlight detection method with respect to the construction gate scenario.

In the testing, three videos represented different illumination conditions (i.e., morning, noon, and evening). It was found that the performance of the proposed method is stable (around 90% of F1 score) in dealing with different illumination conditions. In contrast, the baseline method achieved F1 scores of 82%, 79.1%, and 69%, respectively, for the three test videos. The baseline method was shown to be less effective in dealing with the night-time illumination condition (Gate-Video3) because the feature-based highlight detection method was sensitive to illumination variations. The proposed method adopted object keyframes by tracking construction machines from videos, while the machine tracking module was built upon deep learning object detection. Therefore, the proposed method was found to be more robust than the feature-based method in detecting construction video highlights.

Table 4
Construction rules applied to construction gate case.

# of rule	Rule type	Machine/s of interest	Working zone	d_1 (pixel)	d_2 (pixel)
1	Working zone rule	Dump trucks, concrete mixer trucks, dozers, and cars	[(0,600), (1920,600), (0,1080), (1920,1080)]	NA	NA
2	Working status rule	Dump trucks, concrete mixer trucks, dozers, and cars	[(0,600), (1920,600), (0,1080), (1920,1080)]	60	10

4.2.4. Video highlights for construction gate control

In construction management, gate control is a critical factor in achieving project success. Construction machines should access the gate at the scheduled time to complete their construction tasks, and the timestamp of machines accessing the gate should be recorded. Construction video highlights can serve the gate control purpose by providing video records and corresponding time stamp. Table 6 shows the actual number of instances of machines accessing the gate in the raw video, the number of instances of machines accessing the gate contained in the video highlights, and the accuracy of the three test videos. In case 1, the three test videos showed 48 records of machines accessing the gate, while 45 access records were found to be contained in the detected video highlights, resulting in an accuracy of 93.8%. This result indicates that the video highlight detection method is reasonably effective for construction gate control, and that the generated video highlights can be useful as a form of project documentation for future reference.

4.3. Case study 2: earthmoving

Case study 2 focused on earthmoving, where the proposed method was tested on video footage of an excavator working with several dump trucks. Earthmoving refers to a range of activities that involve excavating soil or rock and moving it to another part of the site, fundamental activities in all types of construction (e.g., residential building, roads, tunnels).

4.3.1. Experimental setup

In case 2, the proposed method was tested on a 40-min earthmoving video with a resolution of 1280 × 720 and a frame rate of 30 fps. In the video, a Volve EC210BLC excavator (bucket payload of 2.1 loose cubic yards (LCY)) works with several dump trucks in an outdoor construction environment and completes several earthmoving cycles. In each cycle, the excavator digs soil and loads it into a dump truck. After the dump truck is fully loaded, it moves away and another dump truck approaches the excavator for the next cycle.

The earthmoving video footage was manually annotated to obtain the ground truth video highlights by following the same procedure described in reference to the construction gate case (i.e., annotators were required to find the video clips of the excavator loading the dump truck, a change in status of the excavator, or any clips that may be of interest for construction management purposes). Through this process, 115 video clips were identified as video highlights in this case study. As with the other case, the feature-based highlight detection method was adopted as the baseline method to test the earthmoving video footage. The configuration of the baseline method was the same as in case 1.

In the earthmoving case, the working zone rule, working status rule, and working interaction rule were applied for detecting object keyframes. The details of these rules are summarized in Table 7. It should be noted that the working zone rule and working status rule only target the excavator, since the excavator is the major construction machine in this

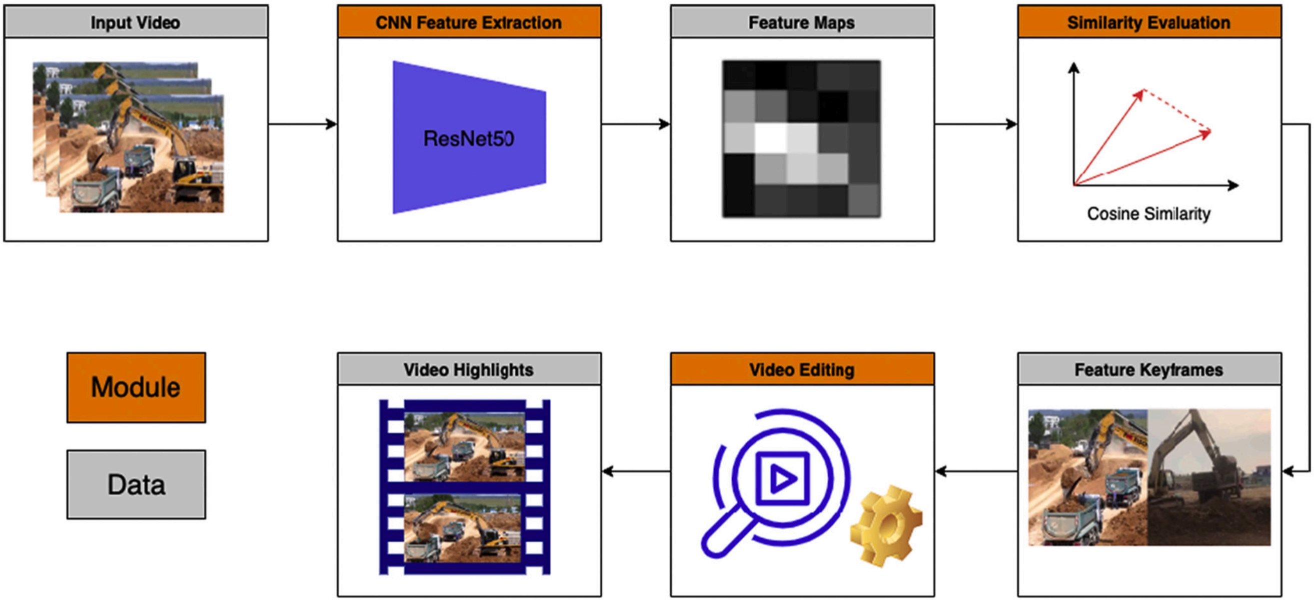


Fig. 8. The overview of the baseline method.

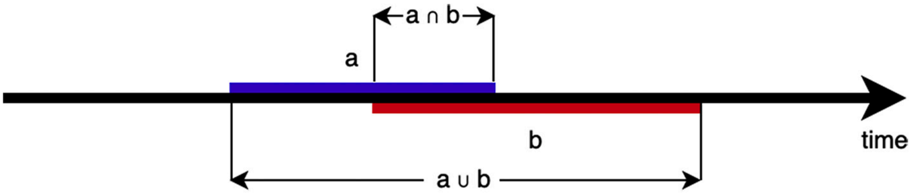


Fig. 9. Illustration of the computation of TIoU.

Table 5
Experimental results of proposed method in construction gate case.

		Precision	Recall	F1 score	Correct highlights detected
Gate-Video1	Proposed method	90%	90%	90%	18
	Baseline method	84.2%	80%	82%	16
Gate-Video2	Proposed method	86.4%	100%	92.7%	19
	Baseline method	70.8%	89.5%	79.1%	17
Gate-Video3	Proposed method	86.7%	92.9%	89.7%	13
	Baseline method	66.7%	71.4%	69%	10
Average	Proposed method	87.7%	94.3%	90.8%	16.7
	Baseline method	73.9%	80.3%	76.7%	14.3

case and it governs the productivity of the whole crew. The working interaction rule, meanwhile, focuses on cases of overlap between the excavator and dump truck.

4.3.2. *Experimental results*

Table 8 shows the experimental results of the proposed method and the baseline method in terms of precision, recall, F1 score, and the number of correct highlights. The proposed method detected 111 video highlights, 104 of them being correct highlights, achieving a precision of 93.7%, recall of 90.4%, and F1 score of 92.0%. Meanwhile, the baseline method achieved a precision of 70.8%, recall of 63.5%, and F1 score of

Table 6
Summary of machines accessing the gate.

	No. of machine accesses in original video	No. of machine accesses in detected video highlights	Accuracy
Gate-Video1	18	16	88.9%
Gate-Video2	18	18	100.0%
Gate-Video3	12	11	91.7%
Sum	48	45	93.8%

Table 7
Construction rules applied to earthmoving case.

# of rule	Rule type	Machine/s of interest	Working zone	d_1 (pixel)	d_2 (pixel)	a
1	Working zone rule	Excavator	[(0,215), (1045,215), (0,720), (1045,720)]	NA	NA	NA
2	Working status rule	Excavator	[(0,215), (1045,215), (0,720), (1045,720)]	20	10	NA
3	Working interaction rule	Excavator and dump truck	[(0,215), (1045,215), (0,720), (1045,720)]	NA	NA	0.1

Table 8
Experimental results of proposed method in earthmoving case.

		Precision	Recall	F1 score	Correct highlights detected
Earthmoving-Video1	Proposed method	93.7%	90.4%	92.0%	104
	Baseline method	84.2%	63.5%	82.0%	73

72.4%. As can be seen, the proposed method outperformed the baseline feature-based method by a margin of 22.9% with respect to precision, 16.9% on recall, and 19.6% in terms of F1 score for the earthmoving case. It is also worth noting that, although the earthmoving case contains more extensive video highlights than the construction gate case, the proposed method achieved similar performance for both cases, underscoring the ability of the proposed method to deal with different construction scenarios.

4.3.3. Video highlights for productivity analysis

In the earthmoving case, the detected highlights were found to contain meaningful video clips of loading activities that would be useful for productivity analysis. As mentioned, in the earthmoving cycle, the excavator digs soil into the bucket and then loads it into a dump truck. Once the dump truck is fully loaded, it moves away and another dump truck approaches the excavator for the next cycle. As such, the number of cycles is equal to the number of loading activities, such that the excavator productivity can be calculated as per Eq. 13, where the bucket payload per cycle is given by the excavator manufacturer (2.1 LCY, in this case).

$$Productivity = \frac{number\ of\ cycles}{time\ (hr)} \times \frac{bucket\ payload}{cycle}\ (LCY) \qquad (13)$$

In Earthmoving-Video1, the excavator has completed 99 work cycles in 40 min and the ground truth productivity is 311.85LCY/hr. By manually analysis the video highlights detected by the proposed method, the author of the present study found 93 video clips of the loading activity. In other words, if the video highlights are used for advanced vision-based method for productivity analysis, the analyzed productivity of Earthmoving-Video1 can reach 292.95LCY/hr. The accuracy of the productivity analysis, then, is 93.9%, which means 93.9% of the relevant productivity information can be retrieved from the detected video highlights without browsing the original construction videos.

5. Discussions

The experimental results indicate that the proposed method can successfully produce video highlights from construction videos for the purpose of reducing manual inspection efforts and digital storage requirements. The research findings and challenges identified in analyzing the test results are discussed below.

- The proposed method exhibited better performance than the feature-based method for detection of construction video highlights. In experiments, the proposed method has achieved an average precision of 89.2%, recall of 93.3%, and F1 score of 91.1% for two case studies, respectively (4 videos in total), while the baseline method has achieved the average precision of 76.5%, recall of 76.1%, and F1 score of 78.0%. The proposed method outperforms the baseline method over 10.0% on three evaluation metrics. The proposed method also achieved close performance when compared with the state-of-the-art method F-DES (fast and deep event summarization) [26] from computer vision. The F-DES implemented the CNN feature extractor and cosine similarity, which has achieved the average precision of 92.3%, recall of 88.3%, and 89.9% of F1 score in three experiments. The F-DES has obtained 3.1% higher results on precision than the proposed method, while the proposed method outperforms 5.0% on recall and 1.2% on F1-score than F-DES. Technically, the proposed method achieved robust performance for two reasons: (1) adopting pre-defined construction rules (i.e., working zone, working status, and working interaction) to detect object keyframes by analyzing machine trajectories. As such, the proposed method explores the context information from construction videos and becomes more robust; and (2) employing ResNet50 to detect feature keyframes to describe scene changes in construction videos. The feature keyframes efficiently represent the video clips that have no construction machines, while improve the precision of the proposed method.
- Reducing the amount of construction video footage is a crucial benefit of applying video highlight detection in construction. In this regard, Table 9 provides a comparison of the original raw video and the detected video highlights in terms of duration and storage size in reference to the two case studies. The average size of the original videos is 635.5 MB. After implementing with the proposed highlight detection method, the average size is reduced to 34.2 MB, a reduction in storage size requirement of approximately 94.6%. The average duration of video highlights is 2.77 min, while the original videos average 55 min in duration. The results indicate that the video highlights generated represent a more watchable synopsis of the raw video, meaning that the use of this method can reduce the amount of effort required in order to maintain construction video documentation.

- Compare to the baseline method, the proposed method is less sensitive to illumination changes, as demonstrated in the construction gate case. Most construction sites are outdoors, and as such illumination changes are frequent in construction video footage. Feature-based highlight detection methods are prone to errantly detect frames that contain significant illumination variations as keyframes, decreasing the accuracy of the highlight detection. In contrast, the proposed method shows stable performance in dealing with illumination changes because it adopts the machine tracking module for object keyframe selection. The CMT tracking method, built upon YOLO-v3 object detection, shows excellent performance in tracking machine trajectories under illumination changes. In this respect, the proposed method exhibits reliable performance even in challenging construction scenarios. To be noted, replacing the backbone object detector YOLO-v3 with more robust methods (e.g. ResNet-based detectors) will provide better results for object keyframe detection, and eventually improve the performance of video highlights detection.
- Compared with feature-based methods, the proposed method has better interpretability and flexibility because it integrates object keyframe selection with feature keyframe selection. The outputs of the proposed method include not only video highlights, but also the intuitive interpretation of selection rationale, such as a machine entering or leaving the frame. This information is beneficial for project management in terms of gate control and productivity

Table 9
Duration and storage size of video highlights in construction gate case.

	Original video		Detected video highlights	
	Duration (minutes)	Storage size (MB)	Duration (minutes)	Storage size (MB)
Gate-Video1	60	713.1	1.33	35.6
Gate-Video2	60	713.5	1.30	39.3
Gate-Video3	60	713.3	1.15	27.8
Earthmoving-Video1	40	402.1	7.30	72.5
Average	55	635.5	2.77	34.2

analysis, as illustrated in the case studies. Furthermore, in the proposed method, the construction rules can be flexibly customized based on the particular needs of a given construction project. For example, the proposed method can generate video highlights that relate only to a specific construction machine (e.g., dump truck), or movement (e.g., machine leaving the site); this is not possible using feature-based highlight detection methods. The proposed method demonstrates the feasibility of rule-based highlights detection methods in construction scenarios.

- The proposed method has three limitations that need to be investigated in the future. First, the parameters of pre-defined construction rules need to be manually set up, and the parameters in one scenario are not typically generically applicable to other scenarios. Second, the performance of the proposed method may be affected if the tracking method fails occasionally because of heavy occlusions, motion blurs, and so forth. It should be noted in this regard that object detection and tracking methods are developing rapidly within the computer vision field. Third, the processing speed of the proposed method should be enhanced. For instance, for the construction gate case described in this paper, it took approximately 2 h to process the one-hour video, and most of the computational resources were dedicated to the CNN feature extraction module.

6. Conclusions and future works

An effective and efficient method for converting construction video footage into concise video data is in high demand in today's construction industry. This paper proposes a novel vision-based method to generate video highlights from construction videos. The proposed method consists of five modules: machine tracking, rule-based keyframe selection, CNN feature extraction, similarity evaluation, and video editing. Two case studies were conducted to validate the performance of the proposed method using construction gate and earthmoving video footage. The proposed method was found to achieve average precision of 89.2% and average recall of 93.3%, outperforming the feature-based highlight detection method. The proposed method can be integrated into several advanced applications that may potentially benefit construction management, including: (1) auto-generating reports from lengthy construction videos; (2) building a query system that searches for clips of interest in the video footage; and (3) quantitatively analyzing construction productivity based on video highlights.

The contributions of this research are three-fold. First, this research has proposed a novel method to detect video highlights from construction videos, while the proposed method outperforms the baseline method over 10% on robustness and precision. Second, three construction rules have been proposed for object keyframes detection including the working zone rule, working status rule, and working interaction rule. By integrating these rules, the detected video highlights are interpretable and flexible, meaning that the resultant construction videos are searchable, filterable, and manageable. Third, the proposed method is shown its feasibility of eliminating over 90% of storage space while retaining most of the useful information for construction video documentation.

Future works will focus on developing an automated process to set up the parameters of construction rules by using machine learning techniques. Moreover, more robust object detection methods (e.g., ResNet-based detectors) will be investigated in the future to replace the YOLO-v3 employed in this research to achieve better performance of construction video highlights detection. Currently, the proposed method simply adopts the cosine similarity metric for feature keyframes detection, while the SSIM (structural similarity index measure) metric will be implemented in the proposed method to replace the cosine similarity in the future. Finally, the parallel coding strategy will be implemented in the future to improve the processing speed of the proposed method. Replacing the ResNet50 feature extractor with small neural networks is another future work to accelerate the processing speed.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Dr. Meng-Han Tsai, from the National Taiwan University of Science and Technology, for kindly providing the test videos for case study 1, and Mr. Chen-Hsuan Wang, Mr. Tzi-Hao Huang, Mr. Leyan Zhong, Mr. Guanlin Meng, and Mr. Menglin Yu for participating in experiments.

References

[1] J.S. Bohn, J. Teizer, Benefits and barriers of construction project monitoring using high-resolution automated cameras, J. Constr. Eng. Manag. 136 (2010) 632–640, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000164.

[2] B. Xiao, Q. Lin, Y. Chen, A vision-based method for automatic tracking of construction machines at nighttime based on deep learning illumination enhancement, Autom. Constr. 127 (2021) 103721, https://doi.org/10.1016/j.autcon.2021.103721.

[3] J. Song, C.T. Haas, C.H. Caldas, Tracking the location of materials on construction job sites, J. Constr. Eng. Manag. 132 (2006) 911–918, https://doi.org/10.1061/(ASCE)0733-9364(2006)132:9(911).

[4] S. Han, S. Lee, A vision-based motion capture and recognition framework for behavior-based safety management, Autom. Constr. 35 (2013) 131–141, https://doi.org/10.1016/j.autcon.2013.05.001.

[5] B. Xiao, Z. Zhu, Two-dimensional visual tracking in construction scenarios: A comparative study, J. Comput. Civ. Eng. 32 (2018), 04018006, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000738.

[6] L. Chen, Y. Wang, M.-F.F. Siu, Detecting semantic regions of construction site images by transfer learning and saliency computation, Autom. Constr. 114 (2020) 103185, https://doi.org/10.1016/j.autcon.2020.103185.

[7] T. Hussain, K. Muhammad, W. Ding, J. Lloret, S.W. Baik, V.H.C. de Albuquerque, A comprehensive survey of multi-view video summarization, Pattern Recogn. 109 (2021) 107567, https://doi.org/10.1016/j.patcog.2020.107567.

[8] Y.-L. Lin, V.I. Morariu, W. Hsu, Summarizing while recording: Context-based highlight detection for egocentric videos, in: International Conference on Computer Vision, IEEE, 2015, pp. 443–451, https://doi.org/10.1109/ICCVW.2015.65.

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

[10] M.-W. Park, A. Makhmalbaf, I. Brilakis, Comparative study of vision tracking methods for tracking of construction site resources, Autom. Constr. 20 (2011) 905–915, https://doi.org/10.1016/j.autcon.2011.03.007.

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

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

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

[14] A. Heydarian, M. Memarzadeh, M. Golparvar-Fard, Automated benchmarking and monitoring of an earthmoving operation's carbon footprint using video cameras and a greenhouse gas estimation model, in: International Conference on Computing in Civil Engineering, ASCE, 2012, pp. 509–516, https://doi.org/10.1061/9780784412343.0064.

[15] W. Zhou, J. Whyte, R. Sacks, Construction safety and digital design: A review, Autom. Constr. 22 (2012) 102–111, https://doi.org/10.1016/j.autcon.2011.07.005.

[16] K. Kumar, D.D. Shrimankar, N. Singh, Event BAGGING: A novel event summarization approach in multiview surveillance videos, in: International Conference on Innovations in Electronics, Signal Processing and Communication, IEEE, 2017, pp. 106–111, https://doi.org/10.1109/IESPC.2017.8071874.

[17] M. Merler, K.-N.C. Mac, D. Joshi, Q.-B. Nguyen, S. Hammer, J. Kent, J. Xiong, M. N. Do, J.R. Smith, R.S. Feris, Automatic curation of sports highlights using multimodal excitement features, IEEE Trans. Multi. 21 (2019) 1147–1160, https://doi.org/10.1109/TMM.2018.2876046.

[18] L. Wang, D. Liu, R. Puri, D.N. Metaxas, Learning trailer moments in full-length movies. http://arxiv.org/abs/2008.08502, 2020.

[19] K. Kumar, D.D. Shrimankar, Deep event learning boosT-up approach: DELTA, Multimed. Tools Appl. 77 (2018) 26635–26655, https://doi.org/10.1007/s11042-018-5882-z.

[20] K. Kumar, D.D. Shrimankar, N. Singh, SOMES: An efficient SOM technique for event summarization in multi-view surveillance videos, in: Recent Findings in

Intelligent Computing Techniques, Advances in Intelligent Systems and Computing, Springer, 2018, pp. 383–389, https://doi.org/10.1007/978-981-10-8633-5_38.

[21] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, B.E. Ionescu, Video summarization from spatio-temporal features, in: 2nd ACM workshop on Video summarization - TVS '08, ACM Press, 2008, pp. 144–148, https://doi.org/10.1145/1463563.1463590.

[22] G. Liu, X. Wen, W. Zheng, P. He, Shot boundary detection and keyframe extraction based on scale invariant feature transform, in: Eighth International Conference on Computer and Information Science, IEEE, 2009, pp. 1126–1130, https://doi.org/10.1109/ICIS.2009.124.

[23] B. Mahasseni, M. Lam, S. Todorovic, Unsupervised video summarization with adversarial LSTM networks, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 2982–2991, https://doi.org/10.1109/CVPR.2017.318.

[24] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S.W. Baik, V.H.C. de Albuquerque, Cloud-assisted multiview video summarization using CNN and bidirectional LSTM, IEEE Trans. Ind. Inform. 16 (2020) 77–86, https://doi.org/10.1109/TII.2019.2929228.

[25] T. Hussain, K. Muhammad, J. Del Ser, S.W. Baik, V.H.C. de Albuquerque, Intelligent embedded vision for summarization of multiview videos in IIoT, IEEE Trans. Ind. Inform. 16 (2020) 2592–2602, https://doi.org/10.1109/TII.2019.2937905.

[26] K. Kumar, D.D. Shrimankar, F-DES: fast and deep event summarization, IEEE Trans. Multi. 20 (2018) 323–334, https://doi.org/10.1109/TMM.2017.2741423.

[27] K. Kumar, Text query based summarized event searching interface system using deep learning over cloud, Multimed. Tools Appl. 80 (2021) 11079–11094, https://doi.org/10.1007/s11042-020-10157-4.

[28] B.T. Truong, S. Venkatesh, Video Abstraction, ACM Transactions on Multimedia Computing, Communications, and Applications 3, 2007, p. 3, https://doi.org/10.1145/1198302.1198305.

[29] R.E. Schapire, Explaining AdaBoost, in: Empirical Inference, Springer, 2013, pp. 37–52, https://doi.org/10.1007/978-3-642-41136-6_5.

[30] P. Mundur, Y. Rao, Y. Yesha, Keyframe-based video summarization using Delaunay clustering, Int. J. Digit. Libr. 6 (2006) 219–232, https://doi.org/10.1007/s00799-005-0129-9.

[31] K. Kumar, EVS-DK: Event video skimming using deep keyframe, J. Vis. Commun. Image Represent. 58 (2019) 345–352, https://doi.org/10.1016/j.jvcir.2018.12.009.

[32] K. Kumar, D.D. Shrimankar, ESUMM: event summarization on scale-free networks, IETE Tech. Rev. 36 (2019) 265–274, https://doi.org/10.1080/02564602.2018.1454347.

[33] K. Muhammad, T. Hussain, S.W. Baik, Efficient CNN based summarization of surveillance videos for resource-constrained devices, Pattern Recogn. Lett. 130 (2020) 370–375, https://doi.org/10.1016/j.patrec.2018.08.003.

[34] Z. Lu, K. Grauman, Story-driven summarization for egocentric video, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2013, pp. 2714–2721, https://doi.org/10.1109/CVPR.2013.350.

[35] K. Zhang, W.-L. Chao, F. Sha, K. Grauman, Video summarization with long short-term memory, in: European Conference on Computer Vision, Springer, 2016, pp. 766–782, https://doi.org/10.1007/978-3-319-46478-7_47.

[36] M. Gygli, H. Grabner, H. Riemenschneider, L. Van Gool, Creating summaries from user videos, in: European Conference on Computer Vision, Springer, 2014, pp. 505–520, https://doi.org/10.1007/978-3-319-10584-0_33.

[37] Song Yale, J. Vallmitjana, A. Stent, A. Jaimes, T.V. Sum, Summarizing web videos using titles, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 5179–5187, https://doi.org/10.1109/CVPR.2015.7299154.

[38] Open Video Project. https://open-video.org, 2021.

[39] S.E.F. de Avila, A.P.B. Lopes, A. da Luz, A. de Albuquerque Araújo, VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recogn. Lett. 32 (2011) 56–68, https://doi.org/10.1016/j.patrec.2010.08.004.

[40] S.K. Kuanar, K.B. Ranga, A.S. Chowdhury, Multi-view video summarization using bipartite matching constrained optimum-path forest clustering, IEEE Trans. Multi. 17 (2015) 1166–1173, https://doi.org/10.1109/TMM.2015.2443558.

[41] Y. Jiao, Z. Li, S. Huang, X. Yang, B. Liu, T. Zhang, Three-dimensional attention-based deep ranking model for video highlight detection, IEEE Trans. Multi. 20 (2018) 2693–2705, https://doi.org/10.1109/TMM.2018.2815998.

[42] M. Sun, A. Farhadi, S. Seitz, Ranking domain-specific highlights by analyzing edited videos, in: European Conference on Computer Vision, Springer, 2014, pp. 787–802, https://doi.org/10.1007/978-3-319-10590-1_51.

[43] B. Xiong, Y. Kalantidis, D. Ghadiyaram, K. Grauman, Less is More: Learning Highlight Detection from Video Duration. http://arxiv.org/abs/1903.00859, 2019.

[44] D. Duque, H. Santos, P. Cortez, The OBSERVER: An intelligent and automated video surveillance system, in: International Conference Image Analysis and Recognition, AIMI, 2006, pp. 898–909, https://doi.org/10.1007/11867586_81.

[45] R. Roy, M. Low, J. Waller, Documentation, standardization and improvement of the construction process in house building, Constr. Manag. Econ. 23 (2005) 57–67, https://doi.org/10.1080/0144619042000287787.

[46] Y.J. Lee, M.W. Park, 3D tracking of multiple onsite workers based on stereo vision, Autom. Constr. 98 (2019) 146–159, https://doi.org/10.1016/j.autcon.2018.11.017.

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

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

[49] B. Xiao, S. Kang, Vision-based method integrating deep learning detection for tracking multiple construction machines, J. Comput. Civ. Eng. 35 (2021), 04020071, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000957.

[50] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[51] R. Jonker, A. Volgenant, A shortest augmenting path algorithm for dense and sparse linear assignment problems, Computing 38 (1987) 325–340, https://doi.org/10.1007/BF02278710.

[52] B. Xiao, S. Kang, Development of an image data set of construction machines for deep learning object detection, J. Comput. Civ. Eng. 35 (2021), 05020005, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000945.

[53] I. Ha, H. Kim, S. Park, H. Kim, Image retrieval using BIM and features from pretrained VGG network for indoor localization, Build. Environ. 140 (2018) 23–31, https://doi.org/10.1016/j.buildenv.2018.05.026.

[54] B. Xiao, K.Y.K. Lam, J. Cui, S.-C. Kang, Perceptions for crane operations, in: Internaltional Conference on Computing in Civil Engineering, ASCE, 2019, pp. 415–421, https://doi.org/10.1061/9780784482438.053.

[55] Z. Kolar, H. Chen, X. Luo, Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images, Autom. Constr. 89 (2018) 58–70, https://doi.org/10.1016/j.autcon.2018.01.003.

[56] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385, 2015.

[57] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252, https://doi.org/10.1007/s11263-015-0816-y.

[58] H.V. Nguyen, L. Bai, Cosine similarity metric learning for face verification, in: Asian Conference on Computer Vision, 2011, pp. 709–720, https://doi.org/10.1007/978-3-642-19309-5_55.