Automation in Construction 120 (2020) 103376

Contents lists available at ScienceDirect

Automation in Construction

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

Towards database-free vision-based monitoring on construction sites: A deep active learning approach

Jinwoo Kim[a,b], Jeongbin Hwang[a], Seokho Chi[a,b,*], JoonOh Seo[c]

[a] Department of Civil and Environmental Engineering, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul, Republic of Korea
[b] Institute of Construction and Environmental Engineering, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul, Republic of Korea
[c] Department of Building and Real Estate, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Construction site
Vision-based monitoring
Database-free
Active learning
Deep learning
Object detection

ABSTRACT

In order to achieve database-free (DB-free) vision-based monitoring on construction sites, this paper proposes a deep active learning approach that automatically evaluates the uncertainty of unlabeled training data, selects the most meaningful-to-learn instances, and eventually trains a deep learning model with the selected data. The proposed approach thus involves three sequential processes: (1) uncertainty evaluation of unlabeled data, (2) training data sampling and user-interactive labeling, and (3) model design and training. Two experiments were performed to validate the proposed method and confirm the positive effects of active learning: one experiment with active learning and the other without active learning (i.e., with random learning). In the experiments, the research team used a total of 17,000 images collected from actual construction sites. To achieve 80% mean Average Precision (mAP) for construction object detection, the random learning method required 720 training images, whereas only 180 images were sufficient when exploiting active learning. Moreover, the active learning could build a deep learning model with the mAP of 93.0%, while that of the random learning approach was limited to 89.1%. These results demonstrate the potential of the proposed method and highlight the considerable positive impacts of uncertainty-based data sampling on the model's performance. This research can improve the practicality of vision-based monitoring on construction sites, and the findings of this study can provide valuable insights and new research directions for construction researchers.

1. Introduction

The 4th edition of "A Guide to the Project Management Body of Knowledge" [1] underlines that "continuous monitoring gives the project management team insight into the health of the project, and identifies any areas that may require special attention." Practitioners and researchers have also acknowledged the importance of construction site monitoring, which is a process of understanding the dynamic and complex natures of construction worksites [2–8]. Continuous monitoring allows project managers to evaluate the operational efficiency of input resources (e.g., direct work rate, hourly production rate), discover potential risk factors that can cause safety accidents (e.g., access to dangerous areas), and understand the current construction progress (e.g., schedule delays). By being aware of the performance and project health of a jobsite, project managers can pay special attention and take proper corrective actions to handle unexpected events, which could adversely affect the project's completion. For example, managers can allocate more dump trucks on site if there are too many loaders waiting for trucks to arrive. Hazardous objects, e.g., holes on worksites, can be identified and removed in advance, and potential accidents can be prevented. This jobsite monitoring and decision-making process can bring an opportunity to enhance on-site performance and enable successful completion of construction projects.

In the past, project managers have directly visited and monitored construction sites manually. However, they have faced difficulties in monitoring dynamic and large-scale jobsites owing to time and cost limitations, and thus many researchers have investigated various automated monitoring systems. One of the most popular systems is an Internet-of-Things-based (IoT-based) approach, which involves attaching electronic sensors to target construction objects, analyzing their physical movements (e.g., locations, speeds, accelerations), and evaluating the operational performance, such as hourly productivity [9] and ergonomic risks [10]. Despite the promising results, there are several practical issues that limit the applications of IoT systems. For example, IoT sensors should be tagged onto every single construction object. This requirement can hinder IoT applications in complex and dynamic

[*] Corresponding author at: Department of Civil and Environmental Engineering, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul, Republic of Korea.
E-mail addresses: jinwoo92@snu.ac.kr (J. Kim), shchi@snu.ac.kr (S. Chi), joonoh.seo@polyu.edu.hk (J. Seo).

https://doi.org/10.1016/j.autcon.2020.103376
Received 18 February 2020; Received in revised form 7 June 2020; Accepted 4 August 2020

construction sites where a significant number of objects exist [11], which means that it would not be possible to attach IoT sensors to all types of construction equipment and tools (e.g., jack hammers, concrete cutting saws) [12]. As an alternative, vision-based construction site monitoring has drawn considerable attention from many practitioners and researchers. It does not require every object to be tagged with camera sensors, and multiple objects can be even tracked at the same time if they appear in a camera's field-of-view. In addition to such technical benefits, the Korean Government has allowed construction companies to include camera installation costs in their safety management budgets since 2016 [13]. This has increased the willingness of construction companies to pay for camera installation at construction sites, and therefore vision-based approaches have become more practical and affordable.

Numerous researchers have investigated the underpinning vision-based algorithms and have customized them to fulfil diverse monitoring purposes, such as productivity measurement [14–21], safety analysis [22–28], and progress measurement [29]. Despite their successful achievements, as most state-of-the-art technologies originate from traditional deep learning algorithms, it is essential to build an extensive and high-quality training image database (DB) [30]. Moreover, the traditional method focuses on the quantity of the training DB, rather than the data quality (i.e., new information included in the data), and thus a significant amount of human effort can be wasted. To address such drawbacks, this paper proposes a deep active learning approach towards DB-free vision-based monitoring on construction sites. DB-free, hereinafter, refers to a concept, in which the purpose is to minimize the volume of training data and the cost of human labeling, while maximizing the monitoring performance. To this end, this research builds upon a deep active learning algorithm that selects the most meaningful-to-learn instances from abundant unlabeled training data, and then learns the selected data first by interacting with human annotators.

This study makes the following contributions. First, this research develops a novel technical framework that can significantly reduce the required number of training images and maximize the performance of vision-based monitoring. Second, the framework can save time and costs of human labeling, enhancing the practicality of vision systems at construction sites. Third, to the authors' knowledge, this is the first attempt to apply deep active learning, which is one of the most prominent emerging pattern-learning algorithms, in the construction domain. Last, the new DB-free approach can provide valuable insights and research directions in the field of vision-based construction monitoring. Following this introduction, this paper reviews existing studies relevant to vision-based construction monitoring. Subsequently, the technical details of the proposed method are explained. Experiments are then conducted using video stream data collected from actual construction sites. The experimental results are analyzed in the next section, and finally, the research contributions and future works are discussed.

2. Literature review

There have been extensive efforts to automatically monitor construction sites using deep learning-based computer vision techniques. Fang et al. [31] and Kim et al. [32], for example, employed a region-based convolutional neural network (R-CNN) to detect various types of construction objects, including workers and equipment. Other researchers have also demonstrated the great performance of deep learning models, including Faster R-CNN, even under harsh analysis conditions, e.g., scale deviations and illumination variations [25,27,31]. These findings have aided the development of computer vision techniques for automated productivity and safety monitoring. Kim et al. [33] fed CNN-based equipment detection results into an earthmoving process simulation model to monitor productivity. Luo et al. [16] built an activity recognition method, composed of CNN and relevant networks, to detect multiple construction resources (e.g., workers, equipment, tools) and interpret their spatial interactions (e.g., size, distance) in order to extract detailed information about the operational efficiency of construction resources. The authors further improved the method by appending Bayesian nonparametric learning to capture workers' activities in far-field surveillance videos [18], and they also proposed a two-stream CNN model for worker activity recognition [14]. Cai et al. [34] developed a two-step long short term memory (LSTM) model to recognize working groups and their activity types. In other studies [35,36], CNN and double-layer LSTM were integrated to learn and analyze the sequential working patterns of heavy equipment. They further improved the deep learning-based method to monitor earthmoving operations from multi-camera views [21]. Bang and Kim [37] also integrated CNN and LSTM models to transform jobsite images to detailed information about the position, status, and quantity of construction resources. Deep learning approaches have also showed promising results in construction safety analysis. Many studies have used CNN-based object detection results to capture safety-related information, such as lapses in wearing personal-protective-equipment [25,27], non-certified operations [26], and access to dangerous zones [38,39]. Kim et al. [40] evaluated the possibility of physical interferences between construction objects using CNN detection results, and Yan et al. [41] proposed a CNN-based method to estimate spatial crowdedness from two-dimensional jobsite images. To overcome the intrinsic shortcomings of CNN models, i.e., frame-by-frame time-independent analysis, Ding et al. [22] proposed a hybrid deep learning model composed of CNN and LSTM to continuously monitor unsafe behaviors of construction workers.

Deep learning algorithms have shown excellent performance on vision-based construction monitoring. However, to train a reliable deep learning model, it is vital to build a high quality and extensive training image DB. This process involves manually labeling target construction objects and/or their operational information, such as object types and locations, on every single image frame. Such manual processes not only require an excessive amount of time and effort, but also have difficulty in representing a wide range of characteristics of different construction objects (e.g., different types and colors of construction equipment), and thus hinder the practical use of vision-based monitoring on construction sites. To solve this problem, researchers have investigated methods for reducing the time and effort required to build a training DB. Liu and Golparvar-Fard [42] examined the feasibility of crowdsourcing techniques, which are an effective way of outsourcing tedious image-labeling tasks to a crowd of non-expert individuals from an online community, such as Amazon Mechanical Turk. As these studies have focused on labeling only workers and their activity types, a recent study by Wang et al. [43] improved the crowdsourcing method to label various safety-rule violations on construction images. However, such crowdsourcing methods still depend on human efforts and cannot reduce the absolute quantity of training data required. In an effort to automate the annotation process, Soltani et al. [44] generated training data from a virtual equipment model and showed promising results in vision-based excavator detection. Braun and Borrmann [45] used building information modeling to annotate types of building elements (e.g., columns, walls, and slabs) and create training images. Despite their valuable efforts, deep learning models that learn from virtual data may have low performance, because real construction images have considerably different visual characteristics, e.g., textures and types of target objects. It would be also difficult to obtain adequate virtual models for every construction object and site. In order to minimize the amount of human labeling required, while also maintaining model performance, this paper proposes a deep active learning approach that selects the most informative data from a set of real construction images, and then teaches the selected data for a deep learning model stage-by-stage. Specifically, the proposed active learning focuses on construction object detection, which is an essential prerequisite for vision-based monitoring.

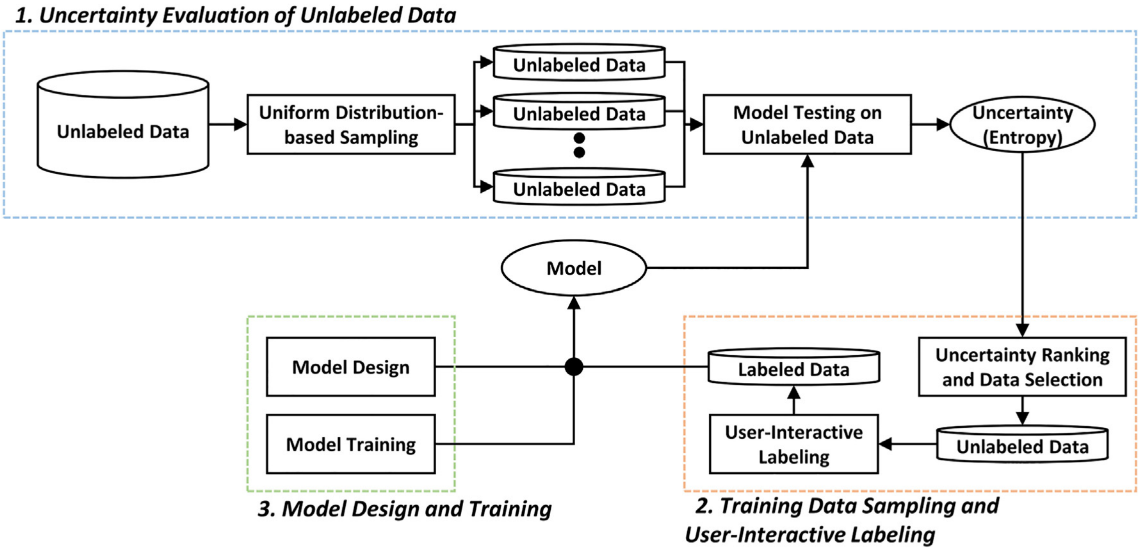

1. Uncertainty Evaluation of Unlabeled Data
Unlabeled Data
Uniform Distribution-based Sampling
Unlabeled Data
Unlabeled Data
Unlabeled Data
Model Testing on Unlabeled Data
Uncertainty (Entropy)
Model
Model Design
Model Training
Labeled Data
Uncertainty Ranking and Data Selection
User-Interactive Labeling
Unlabeled Data
3. Model Design and Training
2. Training Data Sampling and User-Interactive Labeling

Fig. 1. Process flowchart for deep active learning.

3. Proposed approach: deep active learning

Fig. 1 shows the proposed approach, which involves three main processes: (1) uncertainty evaluation of unlabeled data, (2) training data sampling and user-interactive labeling, and (3) model design and training. The details of each process are explained in the following sections.

3.1. Uncertainty evaluation of unlabeled data

The objective of this process is to quantify and evaluate the uncertainty of model prediction for unlabeled training data. First, a sample of unlabeled data is selected through uniform distribution-based random sampling, which means that each sample has an equal probability of being chosen. In this study, 10% of remaining unlabeled data were randomly sampled to reduce computational costs and maintain the model performance. Subsequently, the object detection model trained in the previous step tests multiple image samples, thereby predicting the object type and location (i.e., bounding boxes of each class) of each individual image. In the first training step, a model's parameters can be initialized using the He normal initializer [46], or an open-source pre-trained model by TensorFlow [47] can be used. Based on prediction results, a confidence score for each bounding box can be calculated using the softmax function (Eq. (1.1)), which describes how likely the model thinks each predicted bounding box to be reliable. Finally, the uncertainty for each bounding box is computed as *entropy*, and the uncertainty of each image is determined as the *sum of entropy* of each bounding box (Eq. (1.2)).

$$c_i = \frac{e^{z_i}}{\sum_j e^{z_j}} \, for \; j = 1, 2, ..,J, \tag{1.1}$$

$$E = -\sum_i c_i \log c_i, \tag{1.2}$$

Here, c_i indicates the confidence score of the model's i^{th} prediction (i.e., bounding box) in each image frame; z_i denotes the i^{th} output value from the prior deep learning layer; and E refers to the sum of entropy for an image.

As *entropy* is one of the most popular uncertainty measures that quantify the amount of information required to encode data distribution in the area of information theory [48], *entropy*-based sampling can discover the meaningful-to-learn instances from unlabeled data, enabling to train a more robust object detector. Fig. 2 displays examples of uncertainty evaluation results. As can be seen in the “Low Uncertainty” case, there is no additional new information, because the model already knows which areas are target objects or background. It further means that the images with low uncertainty have a low possibility of improving the model performance due to the lack of new information. For the “High Uncertainty” data in Fig. 2, however, there are many true negatives (i.e., when a target object is not detected) and false positives (i.e., when a detector misclassifies a non-target object or background as a target object). In this case, if human accurately annotates the uncertain-to-predict images (Section 3.2), the model can learn the new additional information, “uncertainties,” (i.e., true negatives and false positives) and evolve more effectively. This uncertainty evaluation is repeated until there is no unlabeled image data left. Eventually, it would be possible to develop a quality-oriented training DB only with high uncertainty images, rather than using existing quantity-oriented approaches.

3.2. Training data sampling and user-interactive labeling

This process selects target training data stage-by-stage based on the results of uncertainty evaluation and asks human to annotate the selected data (i.e., object types and locations within the images) interactively. Specifically, the top 10% of the high uncertainty images are selected for manual labeling, and human annotators perform data labeling using an open-source image labeling software, LabelImg [49]. According to the authors' experiments, the annotation process averagely took 10 s per one image. Fig. 3 displays an example of user-interactive labeling using LabelImg. The annotator can draw bounding boxes on the image by using computer mouse and insert the names of the selected objects. Annotation data [*object_type*, *xmin*, *ymin*, *width*, *height*] are subsequently produced for each image frame. As a result, the previous high quality training DB becomes more informative with accurately labeled object information.

3.3. Model design and training

In this process, the research team designs and trains a deep learning model to detect construction objects using the labeled images. The authors build upon one of the most popular and outstanding object detection models, i.e., Faster Region-proposal CNN (Faster R-CNN).

As depicted in Fig. 4, the Faster R-CNN model comprises three main modules: feature extraction, region proposal, and detection modules. First, a raw red–green–blue (RGB) image is fed into the feature extraction module designed with 13 convolution and five max-pooling

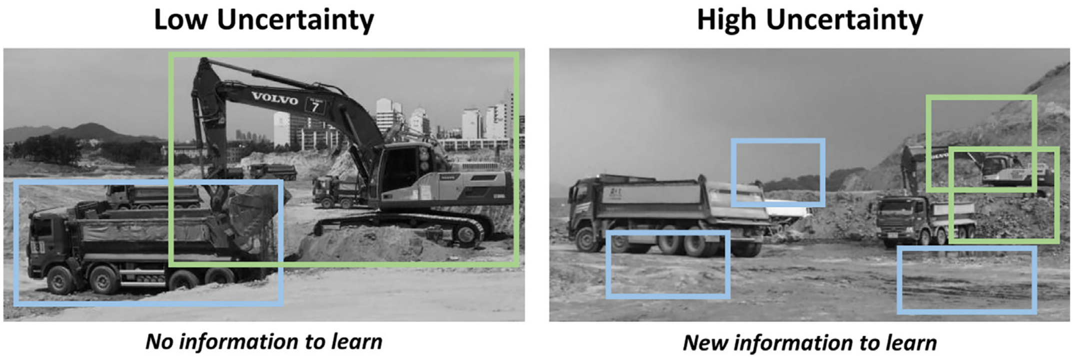


Fig. 2. Examples of low-and high-uncertainty images. In “Low Uncertainty” image, there exists no new information to learn as model can accurately detect excavator and dump truck, whereas “High Uncertainty” image contains new information in terms of true negatives and false positives.

layers. The convolution layers play a key role in extracting visual features from input images, whereas max-pooling layers reduce spatial dimensions of the feature map. The feature map provides key information as to where objects are more likely to exist in an image. Next, the feature map is used as input to the region proposal module, and $n \times n$ spatial windows are slid over the feature map for proposing possible regions of interest (ROIs). To be specific, each window region is processed in three convolution layers, and the results obtained are subsequently fed into two types of fully-connected layers: box-objectness and box-regression. The box-objectness layer computes object and non-object probabilities (i.e., background), whereas the box-regression layer re-corrects proposed ROIs. ROIs classified as objects are then projected onto the feature map with lower dimensions, and each projected region is taken to the detection module comprising box-classification and box-regression layers. Finally, the box-classification layer classifies types of ROI objects and calculates confidence scores using the softmax function (Eq. (1.1)) [50,51]. The box-regression layer predicts two-dimensional coordinates of bounding boxes, [*xmin*, *ymin*, *width*, *height*].

During region proposals, different sizes and aspect ratios of anchor boxes are used to manage scale variations. As default, this research employed 12 anchors in each sliding position. The anchors were characterized by four different scales (32^2, 64^2, 128^2, and 256^2) and three aspect ratios (1:1, 1:2, and 2:1). These default conditions were observed to be sufficient to detect different object types, the appearance and shapes of which continuously changed during the authors' heuristic experiments. Further, to alleviate redundant computation complexity, the number of region proposals pertaining to each image was limited to 100, thereby maintaining consistency with findings of an extant study [50]. In particular, box-regression layers learn two-dimensional coordinate differences between predicted ROIs and the ground truths (Eqs. (2.1)–(2.8)) during the training stage.

$$t_x = (x - x_a)/w_a \tag{2.1}$$

$$t_y = (y - y_a)/h_a \tag{2.2}$$

$$t_w = \log(w/w_a) \tag{2.3}$$

$$t_h = \log(h/h_a) \tag{2.4}$$

$$t_x^* = (x^* - x_a)/w_a \tag{2.5}$$

$$t_y^* = (y^* - y_a)/h_a \tag{2.6}$$

$$t_w^* = \log(w^*/w_a) \tag{2.7}$$

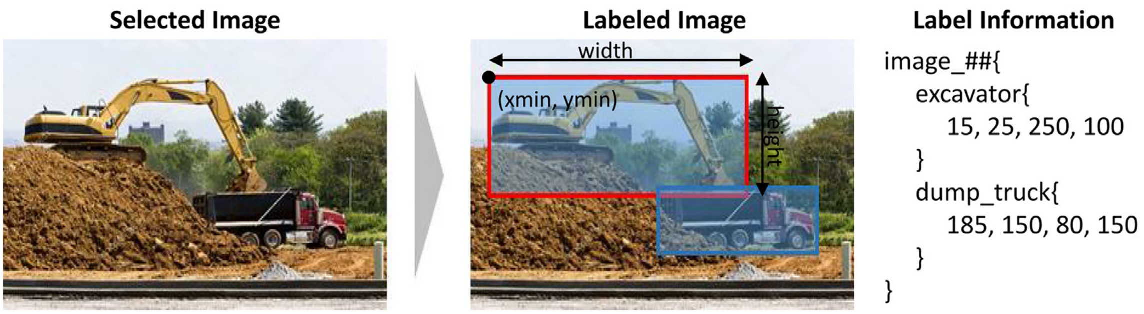
$$t_h^* = \log(h^*/h_a) \tag{2.8}$$

In the above equations, x and y denote center coordinates of the bounding box; w and h denotes the bounding-box width and height; x, x_a, and x^* denote x-coordinates pertaining to the predicted, anchor, and ground truth boxes, respectively (likewise for y); w, w_a, and w^* denote widths of the predicted, anchor, and ground truth boxes, respectively (likewise for h); lastly, $t_{x,\ y,\ w,\ h}$ and $t_{x,\ y,\ w,\ h}{}^*$ denote vectors that represent four parameterized coordinates pertaining to the predicted box and ground truth, respectively.

Using these configurations, the proposed deep learning model was trained using image data selected and labeled, as described in Sections 3.1 and 3.2 (Fig. 1). Stochastic gradient descent was used with a learning rate of 0.0001, a weight decay of 0.00001, a momentum of 0.9, an epoch of 30, and training iterations of 75 per each stage.

4. Experimental results and analysis

To validate the proposed approach, comprehensive experiments were performed using video-stream data collected from four different construction sites. A part of video data was recorded using normal cameras installed at jobsites, whereas others were obtained from an online video-sharing website—YouTube. For the first case, cameras were installed at jobsites according to a systematic camera placement framework developed in the previous research [52,53]. Fig. 5 shows examples of collected video data covering four different types of construction equipment: “excavator,” “dump truck,” “forklift,” and “loader.” The data obtained from various sources facilitated reflecting



```
image_##{
   excavator{
      15, 25, 250, 100
   }
   dump_truck{
      185, 150, 80, 150
   }
}
```

Fig. 3. Example of user-interactive labeling. Annotation data [*object_type, xmin, ymin, width, height*] are generated for each image.

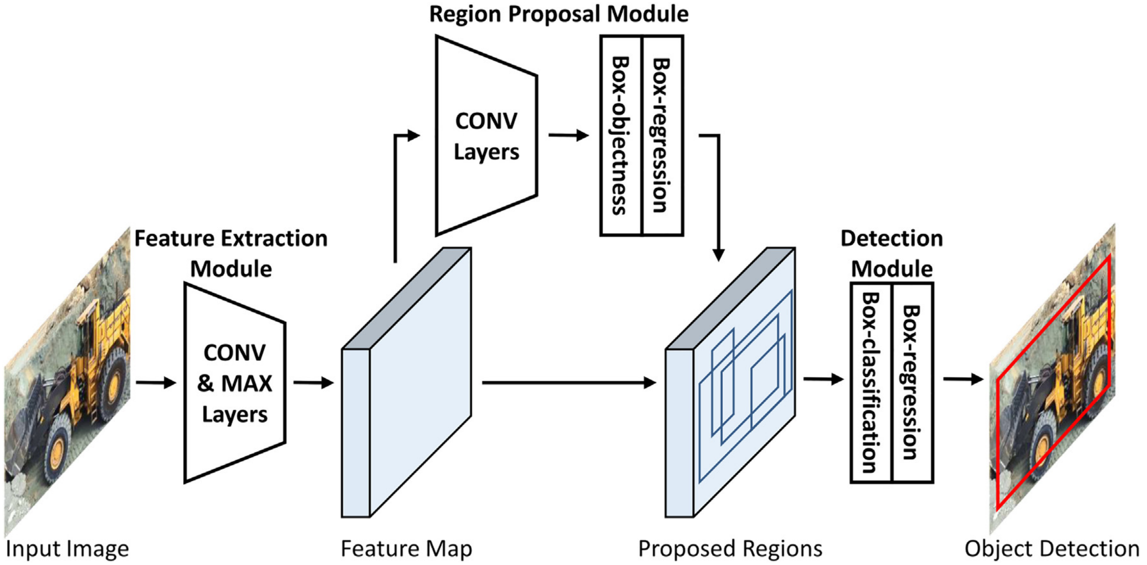


Fig. 4. Working process of Faster R-CNN model that extracts feature maps from input image, proposes possible regions, and detects target objects.

the real-world analytics conditions of occlusions, illuminations, and pose/viewpoint/scale variations. A total of 17,000 image frames were collected with 1280 × 720 resolution and 10 frames per second. The collected data were further split into training and testing data comprising 9400 (55.3%) and 7600 images (44.7%), respectively. Object detection models were trained using two different learning approaches: one with active learning and the other without active learning (i.e., random learning). Random learning is a previously used approach involving random selection of training instances from unlabeled data to develop a training DB. This allowed the authors to confirm both applicability and practicality of the proposed active learning approach as well as determine its beneficial effects.

The performance of the trained models were measured with the mean Average Precision (mAP), which is one of the most popular metrics in vision-based object detection challenges, e.g., PASCAL VOC [54].

4.1. Performance of the proposed approach

Table 1 compares performances of all models considered in this study in terms of mAP values depending on the volume of training data used. The highest performance of the proposed method was approximately 93.0% with a total of 1110 training images. The developed Faster R-CNN model successfully detected various types of construction objects, such as “excavator” (91.3%), “dump truck” (94.7%), “forklift” (95.1%), and “loader” (91.0%). To achieve similar performance, it is general to build a training DB composed of over 10,000 images, as reported in the authors' previous studies: [17,21,35]. Compared to those results, the proposed active learning method could significantly reduce the labeling time from 1800 to 185 min, approximately 90% reduction.

To confirm the expandability of the proposed approach, other types of deep learning models were also examined in this study using active learning. The single shot detector (SSD) [55] and you only look once (YOLO) [56], which are well-known architectures for object detection, were trained. Table 1 summarizes quantitative results obtained by using each model. As can be observed, the Faster R-CNN model demonstrated the best performance among all models with 93% mAP, whereas performances of the SSD and YOLO models were given by mAP values of 92.1% and 89.7%, respectively. Moreover, YOLO showed the highest convergence speed compared to the SSD and Faster R-CNN models. These results are consistent with the existing body of knowledge in computer vision, i.e., complex models usually demonstrate better performance while achieving slow convergence.

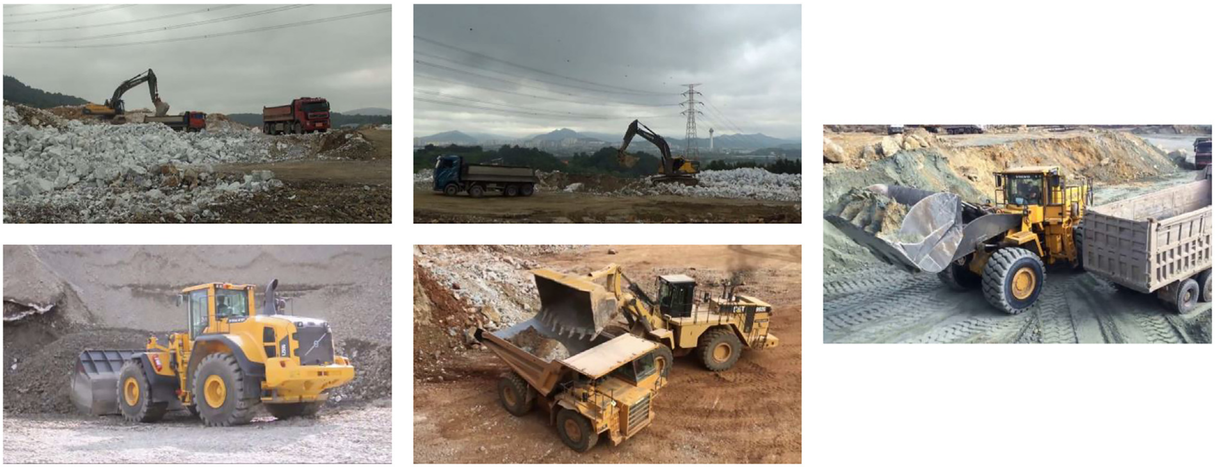


Fig. 5. Visual data samples collected from four different construction sites covering various construction objects and their characteristics, such as illumination and pose/viewpoint/scale variations.

Table 1
Quantitative results in accordance with the number of training images, learning methods, and model architectures.

Number of training images	Faster R-CNN (Active learning)	Faster R-CNN (Random learning)	Single shot detector (Active learning)	You only look once (Active learning)
30	0.406	0.176	0.376	0.369
60	0.432	0.218	0.422	0.432
90	0.570	0.308	0.599	0.589
120	0.612	0.350	0.643	0.633
150	0.701	0.384	0.733	0.720
180	0.803	0.486	0.834	0.744
210	0.779	0.541	0.809	0.741
240	0.905	0.557	0.888	0.852
270	0.869	0.511	0.858	0.827
300	0.827	0.493	0.805	0.796
330	0.877	0.495	0.868	0.836
360	0.908	0.570	0.894	0.873
390	0.924	0.599	0.902	0.863
420	0.907	0.613	0.891	0.869
450	0.921	0.620	0.898	0.897
480	0.881	0.616	0.852	0.841
510	0.918	0.665	0.896	0.857
540	0.871	0.649	0.846	0.824
570	0.896	0.638	0.868	0.846
600	0.914	0.648	0.907	0.873
630	0.909	0.600	0.886	0.868
660	0.900	0.636	0.902	0.859
690	0.879	0.645	0.851	0.830
720	0.912	0.806	0.893	0.861
750	0.895	0.722	0.886	0.844
780	0.919	0.773	0.921	0.882
810	0.906	0.741	0.884	0.842
840	0.893	0.769	0.860	0.846
870	0.891	0.715	0.874	0.849
900	0.895	0.725	0.857	0.843
1050	0.902	0.753	0.877	0.854
1080	0.888	0.742	0.872	0.839
1110	0.930	0.740	0.902	0.878

4.2. Performance with and without active learning

The authors also observed significant impacts of the active learning from the experiments. Fig. 6 shows the learning curves of the proposed and traditional methods, which visualize the models' performances according to the volume of training data used. Under the same conditions (i.e., a set of identical training data and the same prediction model), the performance of the model with the active learning method was always greater than that of the random learning method. From the first training stage with 30 images, the performance difference could be easily recognized. The model trained using active learning was able to detect construction objects with 40.1% mAP, while the performance of the random learning remained in the vicinity of 17.6%. This difference decreased over the entire span of training stages, but did not disappear completely. In detail, the active learning could obtain mAP of 70% with only 150 images (1.6% of the training data), whereas the random learning approach required at least 690 images. To achieve mAP of 80%, 180 and 720 images were required for the active learning and random learning approaches, respectively. In the case of random learning, the highest performance was 89.1%, and a total of 3300 images were required to achieve this value. In contrast, only 240 images were required to achieve the same mAP of 89.1% for the model trained using active learning. Additionally, the performance of the active learning reached the level of over 90% with only 390 training images and keep remained; the best mAP was about 93.0%. These findings indicate that the active learning approach can not only reduce the number of training data required to generate a robust deep learning model but also increase the model's performance.

5. Results and discussion

The proposed method performed well in detecting construction objects from images collected at actual construction sites. The model localized the different types of construction resources, such as "excavator," "dump truck," "forklift," and "loader," with the mAP of 93.0%. Fig. 7 depicts the sequential detection results of the developed model. The proposed method enabled to localize heavy equipment even when the equipment was partially occluded by obstacles (e.g., other nearby dump trucks) or it was not visible from the camera's field-of-view due to dynamic movements. The method was also able to correctly classify construction objects that have diverse visual characteristics (e.g., size, color, and shape) under different camera positions and cluttered backgrounds. These results imply that the proposed model is sufficiently robust to handle a range of image qualities affected by image resolution, presence of occlusions, illuminations, and pose/viewpoint/scale variations.

The results also demonstrated the considerable impacts of active learning on model training. As reported by the experiments, the mAPs over the entire training stages were higher for the active learning approach than for random learning (Fig. 6). It means that the proposed method can build an object recognition model with the same performance, while requiring less training data and reducing the required labeling effort. For example, to generate a model with mAP of over 80%, the required number of training data samples decreased from 720 to 180 images when the active learning was applied. These benefits can be explained by the results of uncertainty evaluation and data sampling (Section 3.1), which is a vital component of active learning. Fig. 8

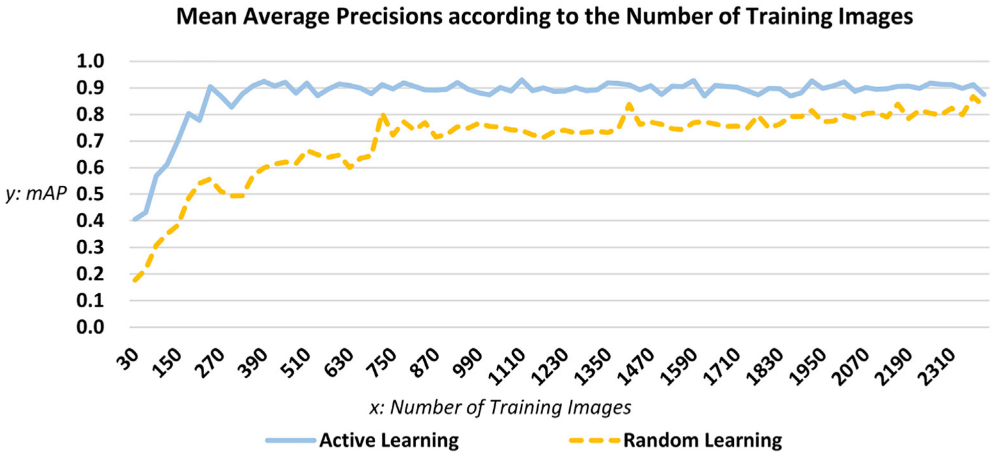


Fig. 6. Learning curves pertaining to active- and random-learning-based models.

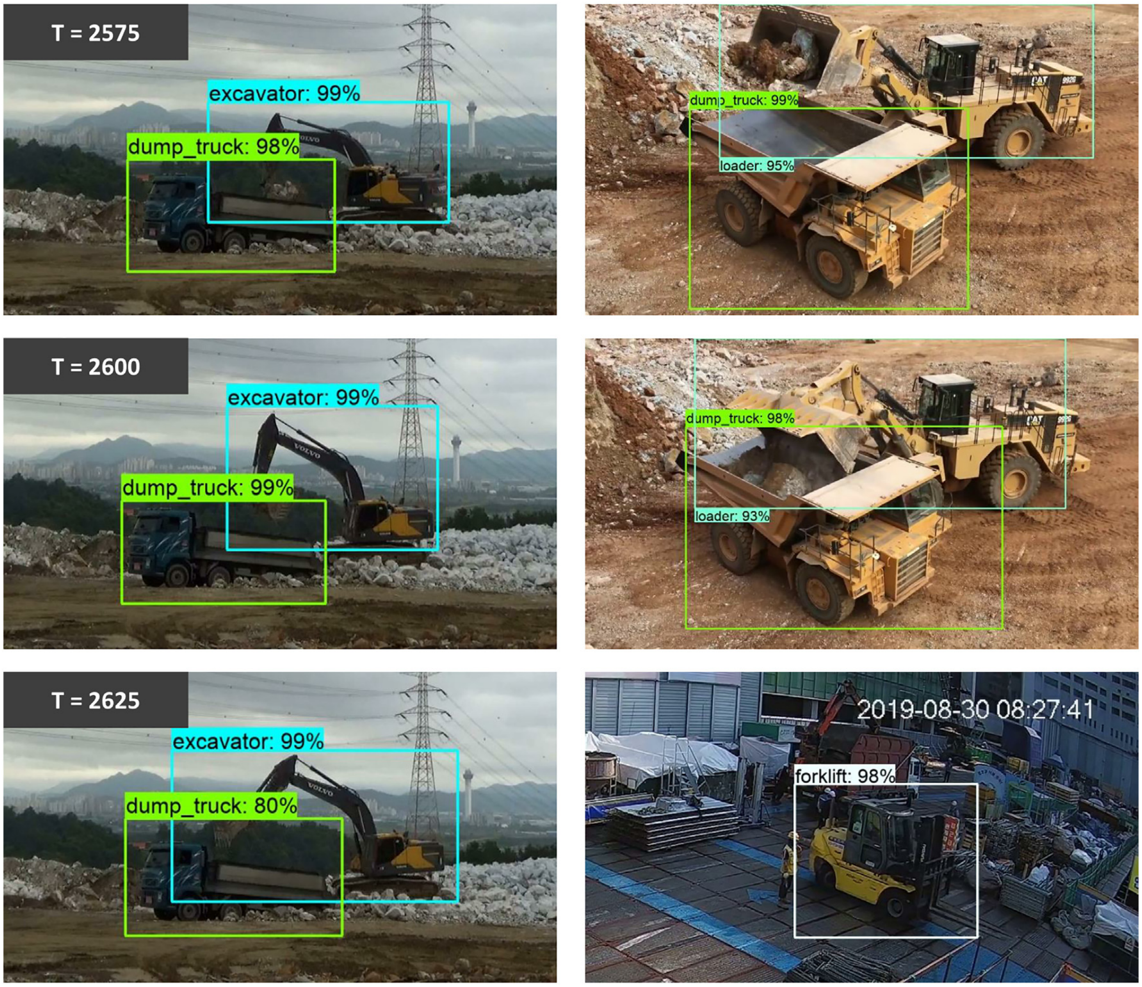


Fig. 7. Experimental results of construction object detection. The developed model can detect different construction objects under different conditions involving occlusions, illuminations, and pose/viewpoint/scale variations.

shows the model's predictions, i.e., bounding boxes and confidence scores, on the images selected as the training data for active and random learning; the images at the left column were sampled based on the uncertainty evaluation (active learning), and the images at the right column were selected through the random sampling (random learning). In the case of active learning, when the first 30 images were investigated, total 25 out of the 30 images had high uncertainty with some false alarms, as marked as "High Uncertainty". This indicates that the model can be enhanced if human performs the user-interactive data labeling (Section 3.2) and corrects the occurred false alarms in the "High Uncertainty" images. The other five images with low uncertainty may not improve the detection performance even if they are used as the training data; it is because the model already knows which areas are target objects or background. In contrast, for the random learning, only six among the 30 images were "High Uncertainty" and the remaining 24 images were "Low Uncertainty". It implies that only six images were meaningful-to-learn and the other ones could not contribute to improving the model performance. Quantitatively, the model's performance was increased by 12.6% for active learning, whereas only 1.6% increment for random learning. It can be concluded that the proposed method has a great ability to sample the meaningful-to-learn instances from abundant unlabeled data, and thereby accelerate the performance improvement.

It was also notable that the model performance was averagely 13.6% higher when using active learning, compared to the random learning. Besides, the active learning ultimately raised the model's final performance by 3.9%, which is approximately 2% greater than the results of the previous studies: [57,58]. It seems that the detector successfully evolved by actively selecting and learning the most uncertain and informative images at every training stage, resulting in the best performance at the end. This can be quantitatively explained by the coefficient of determination R^2—a well-known indicator that explains how stably and sensitively a model evolves through training stages, i.e., model convergence [59]. As shown in Fig. 9, when fitting the derivatives of the learning curves to a logarithmic function, the values of R^2 were approximately 0.171 for active learning and 0.092 for random learning. This implies that when using active learning, a model converges about twice more stable than random learning. Therefore, using the proposed approach can lower the risk of being stuck at a local optimum across all training stages. These findings indicate that the proposed method affords significant advantages in terms of model convergence and overall performance.

6. Conclusions

In working towards DB-free vision-based monitoring on construction sites, this study presented a deep active learning approach that automatically evaluates the uncertainty of training images, selects the most meaningful-to-learn data, and trains a deep learning model using the selected data in a sequential manner. The proposed approach involved three main processes: (1) uncertainty evaluation of unlabeled data, (2) training data sampling and user-interactive labeling, and (3) model design and training. Through the active learning approach, it was possible to minimize the human effort required for data labeling

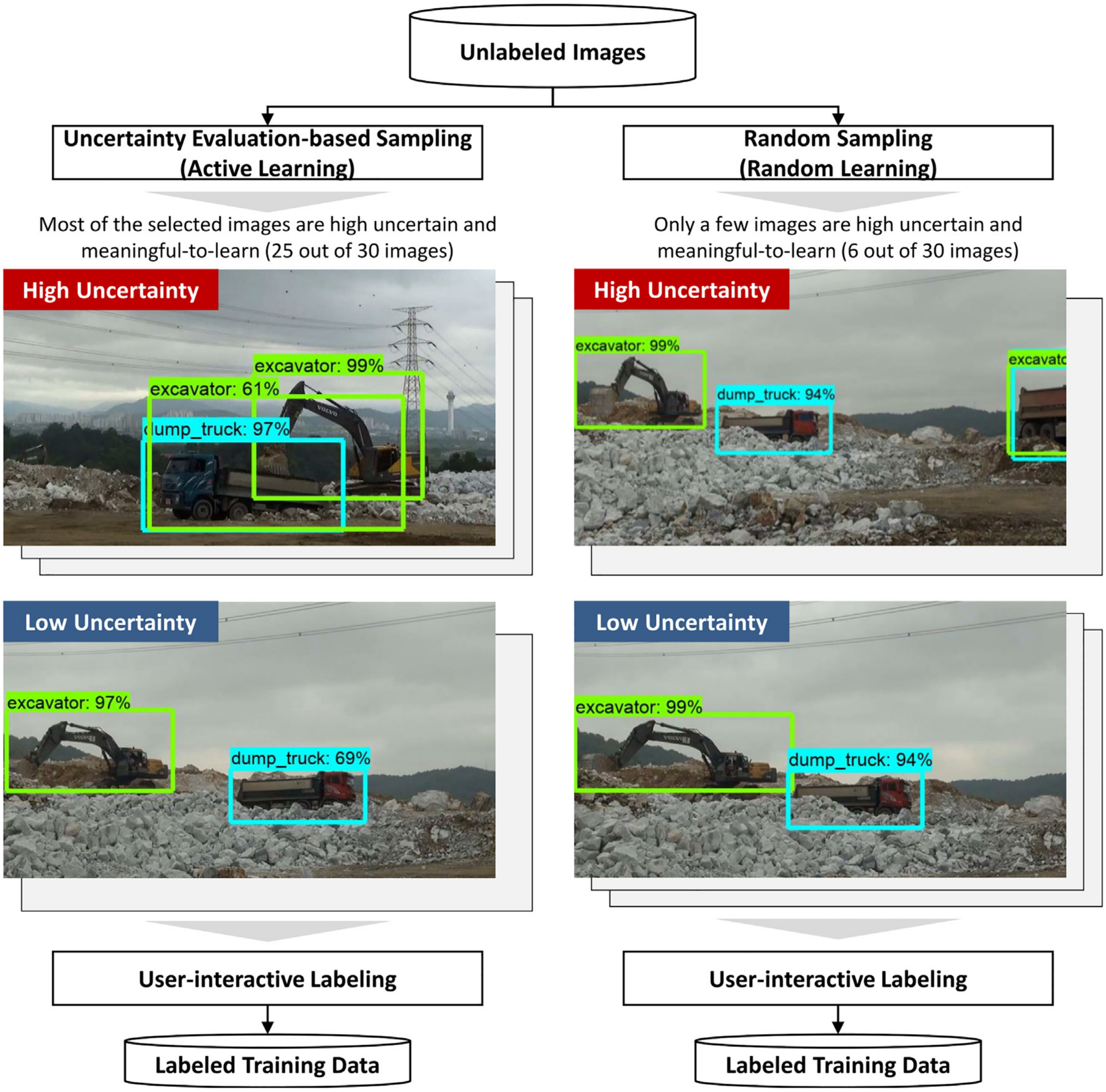


Fig. 8. Examples of images selected for active and random learning.

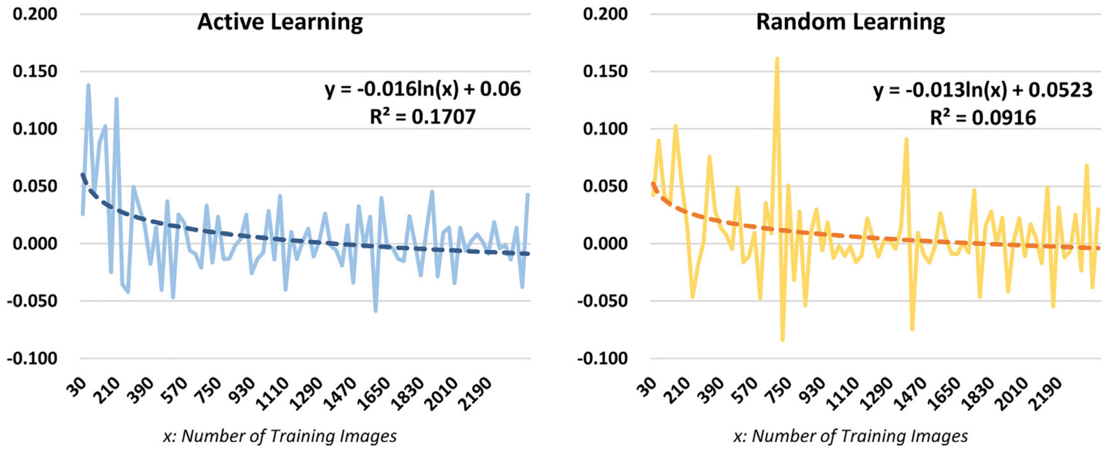


Fig. 9. Learning curve derivatives and their logarithmic fitting. The model converges about twice more stable when using active learning than using random learning.

and to train more powerful object detectors. As reported in the experiments conducted with and without active learning, the number of training images required to obtain mAP values of 70% and 80% were reduced by 540 images, when the active learning approach was used. Furthermore, while the performance of the best model remained at 89.1% in the traditional method (i.e., random learning), the active learning approach enhanced the model's performance up to 93.0%, implying that the optimal solutions at each training stage finally resulted in the near-global optimum. The object detection model also evolved about twice more stably in the authors' experiments with the active learning approach.

Considering the benefits of the proposed method, this study made the following contributions. First, this study documented the development of a novel technical framework that can both significantly reduce the required number of training images and maximize the performance of vision-based monitoring. Second, the framework can save time and costs needed for human labeling, thereby enhancing the practical acceptability of vision systems on construction sites. Third, to the authors' knowledge, this study represented the first attempt to apply deep active learning, which is one of the most prominent and emerging pattern-learning algorithms, in the construction domain. Last, the novel DB-free approach can provide valuable insights and new research directions in the field of not only vision-based construction monitoring but also other research areas, e.g., natural language processing [60], sound recognition [61,63], and wearable sensing [62].

Building on the interesting findings of this study, there are several opportunities for further research. As further efforts to achieve DB-free vision-based construction monitoring, the active learning approach can be integrated with other state-of-the-art technologies. For instance, when applying crowdsourcing labeling techniques [42,43], active learning can recommend informative-to-learn data to crowds (i.e., human annotators) and train a model more robustly and efficiently. Integration with virtual models is another effective way of reducing human effort needed to training data acquisition. One study [44] featured the successful training of an object detection model using images generated from virtual equipment models. The model (learned from virtual images) can be fine-tuned using active learning, and thereby it would be available to further reduce the required volume of training data and human effort, when compared to the learning-from-scratch method (i.e., when training a completely new model). With further achievements, it is believed that DB-free vision-based monitoring can become a reality in future construction projects.

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

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

References

[1] Project Management Institute, A guide to project management body of knowledge (PMBOK guide), 4th ed., Newtown Square, 2018. ISBN 1933890517.

[2] 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.

[3] Y. Ham, K.K. Han, J.J. Lin, M. Golparvar-Fard, Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): a review of related works, Visualization in Engineering. 4 (2016) 1–8, https://doi.org/10.1186/s40327-015-0029-z.

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

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

[6] 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.

[7] B. Zhong, H. Wu, L. Ding, P.E.D. Love, H. Li, H. Luo, L. Jiao, Mapping computer vision research in construction: developments, knowledge gaps and implications for research, Autom. Constr. 107 (2019) 102919, https://doi.org/10.1016/j.autcon.2019.102919.

[8] J. Kim, Automated Construction Site Monitoring: Multi Vision-based Operational Context Analysis for Enhancing Earthmoving Productivity, Seoul National University, 2019, http://hdl.handle.net/10371/161874 , Accessed date: 14 May 2020.

[9] A. Montaser, O. Moselhi, Truck+ for earthmoving operations, Journal of Information Technology in Construction. 19 (2014) 412–433 http://www.itcon.org/2014/25 , Accessed date: 14 May 2020.

[10] N.D. Nath, R. Akhavian, A.H. Behzadan, Ergonomic analysis of construction worker's body postures using wearable mobile sensors, Appl. Ergon. 62 (2017) 107–117, https://doi.org/10.1016/j.apergo.2017.02.007.

[11] M.W. Park, I. Brilakis, Construction worker detection in video frames for initializing vision trackers, Autom. Constr. 28 (2012) 15–25, https://doi.org/10.1016/j.autcon.2012.06.001.

[12] C.F. Cheng, A. Rashidi, M.A. Davenport, D.V. Anderson, Activity analysis of construction equipment using audio signals and support vector machines, Autom. Constr. 81 (2017) 240–253, https://doi.org/10.1016/j.autcon.2017.06.005.

[13] Korea Construction Technology Promotion Act, Enforcement decree article 98 and 99, statutes of the Republic of Korea, http://law.go.kr/법령/건설기술관리법, (2016) , Accessed date: 28 January 2019.

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

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

[16] X. Luo, H. Li, D. Cao, F. Dai, J. Seo, S. Lee, Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks, J. Comput. Civ. Eng. 32 (2018) 04018012, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000756.

[17] J. Kim, S. Chi, J. Seo, Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks, Autom. Constr. 87 (2018) 297–308, https://doi.org/10.1016/J.AUTCON.2017.12.016.

[18] X. Luo, H. Li, X. Yang, Y. Yu, D. Cao, Capturing and understanding workers' activities in far-field surveillance videos with deep action recognition and Bayesian nonparametric learning, Computer-Aided Civil and Infrastructure Engineering. 34 (2019) 333–351, https://doi.org/10.1111/mice.12419.

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

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

[21] J. Kim, S. Chi, Multi-camera vision-based productivity monitoring of earthmoving operations, Autom. Constr. 112 (2020) 103121, https://doi.org/10.1016/j.autcon.2020.103121.

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

[23] S. Han, S. Lee, F. Peña-Mora, Comparative study of motion features for similarity-based modeling and classification of unsafe actions in construction, J. Comput. Civ. Eng. 28 (2014) A4014005, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000339.

[24] J.O. Seo, A. Alwasel, S. Lee, E.M. Abdel-Rahman, C. Haas, A comparative study of in-field motion capture approaches for body kinematics measurement in construction, Robotica. 37 (2019) 928–946, https://doi.org/10.1017/S0263574717000571.

[25] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T.M. Rose, W. An, Detecting non-hardhat-use by a deep learning method from far-field surveillance videos, Autom. Constr. 85 (2018) 1–9, https://doi.org/10.1016/J.AUTCON.2017.09.018.

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

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

[28] H. Kim, K. Kim, H. Kim, Vision-based object-centric safety assessment using fuzzy inference: monitoring struck-by accidents with moving objects, J. Comput. Civ. Eng. 30 (2016) 04015075, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000562.

[29] A. Dimitrov, M. Golparvar-Fard, Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections, Adv. Eng. Inform. 28 (2014) 37–49, https://

doi.org/10.1016/j.aei.2013.11.002.

[30] J. Kim, S. Chi, Adaptive detector and tracker on construction sites using functional integration and online learning, J. Comput. Civ. Eng. 31 (2017) 04017026, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000677.

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

[32] H. Kim, H. Kim, Y.W. Hong, H. Byun, Detecting construction equipment using a region-based fully convolutional network and transfer learning, J. Comput. Civ. Eng. 32 (2018) 04017082, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000731.

[33] H. Kim, S. Bang, H. Jeong, Y. Ham, H. Kim, Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation, Autom. Constr. 92 (2018) 188–198, https://doi.org/10.1016/j.autcon.2018.04.002.

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

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

[36] J. Kim, S. Chi, M. Choi, Sequential pattern learning of visual features and operation cycles for vision-based action recognition of earthmoving excavators, in: Computing in Civil Engineering 2019: Data, Sensing, and Analytics, American Society of Civil Engineers, Atlanta, USA, 2019, pp. 298–304, https://doi.org/10.1061/9780784479247.083.

[37] S. Bang, H. Kim, Context-based information generation for managing UAV-acquired data using image captioning, Autom. Constr. 112 (2020) 103116, https://doi.org/10.1016/j.autcon.2020.103116.

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

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

[40] D. Kim, M. Liu, S. Lee, V.R. Kamat, Remote proximity monitoring between mobile construction resources using camera-mounted UAVs, Autom. Constr. 99 (2019) 168–182, https://doi.org/10.1016/j.autcon.2018.12.014.

[41] X. Yan, H. Zhang, H. Li, Estimating worker-centric 3D spatial crowdedness for construction safety management using a single 2D camera, J. Comput. Civ. Eng. 33 (2019) 04019030, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000844.

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

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

[44] M.M. Soltani, Z. Zhu, A. Hammad, Automated annotation for visual recognition of construction resources using synthetic images, Autom. Constr. 62 (2016) 14–23, https://doi.org/10.1016/j.autcon.2015.10.002.

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

[46] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, Santiago, Chile, 2015, pp. 1026–1034, https://doi.org/10.1109/ICCV.2015.123.

[47] TensorFlow, TensorFlow object detection demo, GitHub. (2019). https://github.com/tensorflow/models/blob/master/research/object_detection/object_detection_tutorial.ipynb (accessed January 9, 2020).

[48] B. Settles, Active learning literature survey, Computer Science Technical Report 1648 (2010), http://burrsettles.com/pub/settles.activelearning.pdf , Accessed date: 14 May 2020.

[49] LabelImg, (2018). https://github.com/tzutalin/labelImg (accessed December 30, 2019).

[50] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, STAT: spatial-temporal attention mechanism for video captioning, IEEE Transactions on Multimedia. 22 (2020) 229–241, https://doi.org/10.1109/TMM.2019.2924576.

[51] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, F. Xu, 3D room layout estimation from a single RGB image, IEEE Transactions on Multimedia. 14 (2020) 1–12, https://doi.org/10.1109/tmm.2020.2967645.

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

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

[54] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2010) 303–338, https://doi.org/10.1007/s11263-009-0275-4.

[55] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, SSD: single shot multibox detector, Lect. Notes Comput. Sci 9905 (2016) 21–37, https://doi.org/10.1007/978-3-319-46448-0_2.

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

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

[58] S. Roy, A. Unmesh, V.P. Namboodiri, Deep active learning for object detection, in: Proceedings of the British Machine Vision Conference 2018 (BMVC 2018), British Machine Vision Association, Newcastle, UK, 2018, pp. 1–12 http://bmvc2018.org/contents/papers/0287.pdf , Accessed date: 14 May 2020.

[59] C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 14 (2020) 1–8, https://doi.org/10.1109/tpami.2020.2975798.

[60] T. Kim, S. Chi, Accident case retrieval and analyses: using natural language processing in the construction industry, J. Constr. Eng. Manag. 145 (2019) 04019004, , https://doi.org/10.1061/(ASCE)CO.1943-7862.0001625.

[61] K. Min, M. Jung, J. Kim, S. Chi, Sound event recognition-based classification model for automated emergency detection in indoor environment, in: Advances in Informatics and Computing in Civil and Construction Engineering, Springer, Chicago, USA, 2018, pp. 529–535, https://doi.org/10.1007/978-3-030-00220-6_63.

[62] G. Lee, B. Choi, H. Jebelli, C.R. Ahn, S. Lee, Wearable biosensor and collective sensing–based approach for detecting older adults' environmental barriers, J. Comput. Civ. Eng. 34 (2020) 04020002, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000879.

[63] J. Kim, K. Min, M. Jung, S. Chi, Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition, Build. Environ. 181 (2020) 107092, https://doi.org/10.1016/j.buildenv.2020.107092.