ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0447
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

3D Visualization and Work Status Analysis of Construction Site Objects

Junghoon Kim[1]*, Insoo Jeong[2], Seungmo Lim[3], Jeongbin Hwang[4], Seokho Chi[5],

[1] *Graduate Student, Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea,* E-mail address: pable01@snu.ac.kr
[2] *Graduate Student, Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea,* E-mail address: scout1120@snu.ac.kr
[3] *Graduate Student, Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea,* E-mail address: lssm818@snu.ac.kr
[4] *Senior Researcher, Institute of Construction and Environmental Engineering (ICEE), Republic of Korea,* E-mail address: jb.hwang@hotmail.com
[5] *Professor, Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea; Adjunct Professor, the Institute of Construction and Environmental Engineering (ICEE), Republic of Korea,* E-mail address: shchi@snu.ac.kr

Abstract: Construction site monitoring is pivotal for overseeing project progress to ensure that projects are completed as planned, within budget, and in compliance with applicable laws and safety standards. Additionally, it seeks to improve operational efficiency for better project execution. To achieve this, many researchers have utilized computer vision technologies to conduct automatic site monitoring and analyze the operational status of equipment. However, most existing studies estimate real-world 3D information (e.g., object tracking, work status analysis) based only on 2D pixel-based information of images. This approach presents a substantial challenge in the dynamic environments of construction sites, necessitating the manual recalibration of analytical rules and thresholds based on the specific placement and the field of view of cameras. To address these challenges, this study introduces a novel method for 3D visualization and status analysis of construction site objects using 3D reconstruction technology. This method enables the analysis of equipment's operational status by acquiring 3D spatial information of equipment from single-camera images, utilizing the Sam-Track model for object segmentation and the One-2-3-45 model for 3D reconstruction. The framework consists of three main processes: (i) single image-based 3D reconstruction, (ii) 3D visualization, and (iii) work status analysis. Experimental results from a construction site video demonstrated the method's feasibility and satisfactory performance, achieving high accuracy in status analysis for excavators (93.33%) and dump trucks (98.33%). This research provides a more consistent method for analyzing working status, making it suitable for practical field applications and offering new directions for research in vision-based 3D information analysis. Future studies will apply this method to longer videos and diverse construction sites, comparing its performance with existing 2D pixel-based methods.

Key words: Operational Productivity, 3D Reconstruction, Computer Vision, Status Analysis, Object Segmentation

1. INTRODUCTION

Construction site monitoring involves tracking the progress of the projects and supervising them to ensure they are completed as planned, within budget, and in compliance with relevant laws and safety regulations. Furthermore, this process is integral to bolstering operational efficiency, thereby facilitating

more effective project execution. Operational productivity, defined as the ratio of production outputs (e.g., the quantity of rebar installed) to input resources (e.g., labor hours on the worksite), is one of the key performance indicators in construction operations [1]. Additionally, it provides critical information for various construction management tasks, including project estimation, financial planning, and schedule development [2]. Enhancing this operational productivity is a primary concern for site managers and directly correlates with the company's benefits, further posing a challenge that the construction industry must address.

In the academic realm, numerous researchers have explored various approaches to enhance construction operational productivity. For instance, computer vision-based automatic monitoring technologies have been utilized to automatically detect and track equipment and workers on-site through CCTV systems installed at construction sites. Furthermore, beyond mere detection and tracking, studies have also been conducted to analyze the work status of each piece of equipment and worker, providing productivity information. These technologies offer site managers meaningful data to improve project productivity and are indeed being implemented in various sites for site management purposes. For instance, by providing information on the working hours and output of each object, it is possible to adjust future work plans accordingly and take action in areas where problems arise. While the application of academic research findings in the actual industry sector is moving in a favorable direction, generating benefits, there still exist limitations to the practical application of these technologies.

Currently, the majority of computer vision-based monitoring technologies conduct analyses based on two-dimensional (2D) pixel information. This 2D pixel-based analysis refers to utilizing the bounding box information (i.e., x and y pixel coordinates within the image) of objects detected in the image. Various rule-based analyses utilize this bounding box data to assess the operational status of the equipment. For example, if the center coordinates of an object's bounding box shift beyond a predefined threshold between frames, the object is categorized as moving; if not, it is deemed stationary. Additionally, if the distance between the bounding boxes of two objects falls below a specified threshold, they are considered to be interacting. The state of work, stop, or movement is then determined based on the objects' Intersection over Union (IoU) value. However, the criteria for determining the work status of objects change depending on the camera placement and the field of view (FOV). For example, if the distance for determining the interaction between equipment were 100 pixels for one camera, this criterion would need to be reduced for shots taken closer to the construction site. Thus, 2D pixel-based analysis requires setting manual criteria for each camera, and these criteria must be modified whenever the work site changes. Furthermore, given the nature of construction sites, where the work environment changes over time, the practical application of this technology faces significant challenges.

To address this issue, researchers are exploring various approaches, including using multi-camera or integrating vision and sensor technologies. This study proposes a method for three-dimensional (3D) visualization and work status analysis of construction site objects using 3D reconstruction technology. The proposed framework consists of three stages: (i) single image-based 3D reconstruction, (ii) 3D visualization, and (iii) work status analysis. Through the proposed framework, it is possible to perform more accurate equipment status analyses using 3D spatial information without the need to adjust settings based on camera viewpoints. Furthermore, it is anticipated that the results can be used to conduct project simulations in 3D space, thereby enhancing overall site productivity.

2. LITERATURE REVIEW

In the early stages, the application of computer vision in the construction sector was primarily directed towards addressing fundamental tasks related to the detection and tracking of resources. For instance, Azar and McCabe [3] identified the off-highway dump trucks by employing the Haar-Histogram of Oriented Gradients (HOG) and Blob-HOG techniques. This approach aimed to identify and track dump trucks precisely, which are critical resources in construction operations. Memarzadeh et al. [4] proposed an equipment detection algorithm using HOG-colors and a Support Vector Machine (SVM) classifier. Ji et al. [5] developed an intelligent site monitoring system that detects dump trucks and excavators using a foreground detection algorithm. Xiao et al. [6] proposed a construction machinery tracker that utilizes YOLOv3 to track multiple construction machines simultaneously.

Furthermore, studies have been carried out on analyzing operational productivity through the use of object detection and tracking algorithms. For example, Gong and Caldas [7] presented a vision-based video interpretation model aimed at automatically transforming videos of construction operations into productivity information. Kim et al. [8] introduced an automatic method for evaluating productivity in

tunnel earthwork processes, which combines vision-based contextual inference and construction process simulations through the use of a convolutional neural network (CNN) model. They determined the operational status by analyzing the changes in distance between the bounding boxes of dump trucks and excavators. Rashid and Louis [9] developed synthetic time-series training data for the purpose of recognizing equipment activities utilizing a recurrent neural network (RNN) architecture based on long short-term memory (LSTM). Applying RNN requires a comprehensive training dataset, a challenge they addressed by augmenting synthetic time-series data. Roberts and Golparvar-Fard [10] also assessed operational productivity by analyzing interactions between dump trucks and excavators using a CNN-based approach. They analyzed the equipment's status by calculating its velocity through the movement of the bounding box centers. Kim et al. [11] proposed a vision-based framework for activity identification, concentrating on the interaction between dump trucks and excavators to classify earthwork tasks and evaluate their work cycles. The suggested framework utilized a rule-based approach for activity analysis, focusing on variations in centroid positions between image frames and the sum of pixel differences between frames.

As mentioned above, in the field of construction vision technology, while numerous studies have focused on site monitoring and productivity, the majority have based their analyses on 2D information to infer actual 3D data. Moreover, to estimate 3D data from 2D information, various rules had to be established, which vary depending on camera placement or training datasets, significantly affecting model performance. Researchers have attempted various approaches to overcome the limitations of these 2D pixel-based estimations. For instance, Park et al. [12] corrected the results of 2D pixel-based estimations by tracking designated objects in 2D video frames and correlating the outcomes across multiple pre-calibrated views using epipolar geometry. Luo et al. [13] developed a method to estimate the 2D poses of key points on construction equipment using stacked hourglass and pyramid network architectures, and as future work, presented an approach for analyzing operational status utilizing this method. While this method employed more detailed information than traditional state analysis based on an object's bounding box data, it still faces the limitation of relying on 2D-pixel information. To address this limitation, Assadzadeh et al. [14] utilized virtual models to generate 3D pose information for excavators and trained the model to estimate the 3D poses of actual site equipment. However, it was limited to extracting poses for individual objects only, and the extracted coordinates were not localized, failing to account for interactions between different pieces of equipment.

3. METHODOLOGY

3.1. Single image-based 3D reconstruction

In this phase, the objective is to create a 3D model of an object via single image-based 3D reconstruction. This process encompasses two principal processes: object segmentation and 3D reconstruction. Initially, the object segmentation process segments the object parts from the input image. For 3D reconstruction, the presence of background elements behind the object of interest in the input image can degrade performance. Therefore, an image cropped to include only the object part, excluding the background, is extracted. In this study, the Sam-Track model was employed for the object segmentation task. Segment Anything Model (SAM) is a foundation model for image segmentation trained on a dataset of 11 million images and 1.1 billion masks, released by the Meta AI research team [15]. Sam-Track is an integrated model combining the SAM model with Decoupling features in Associating Objects with Transformers (DeAOT) for multi-object tracking [16]. By inserting a video and providing text for the desired object, it automatically segments the specified object. In this study, segmentation was performed on excavators and dump trucks, and the results are illustrated in Figure 1.

Figure 1. Example of object segmentation results

Subsequently, the previously obtained cropped image is utilized to reconstruct the object into a 3D model. For the 3D reconstruction process, the One-2-3-45 model was employed. This model generates multiple images from a single image at various angles using a stable diffusion model and then reconstructs these images into 3D using the Neural Radiance Fields (NeRf) model [17]. An example of the 3D reconstruction is shown in Figure 2.

Figure 2. Example of 3D reconstruction

3.2. 3D visualization

In this phase, the objective is to localize the 3D models within a single 3D space, reflecting the object's location in the image. The 3D models generated through the 3D reconstruction process each possess a local coordinate system centered at $(0, 0, 0)$. This is because the segmentation image extracted for 3D model creation from the original image results in the loss of the object's positional information within the original image. Therefore, to place multiple objects in one space and reflect their original positions as observed in the captured footage, a transformation to a global coordinate system is necessary. In this study, Unity was utilized as a tool for 3D localization. Unity is a versatile platform for 3D visualization, virtual reality, and model simulation, among other applications. Within Unity, users can adjust camera settings, facilitating the simulation of actual camera characteristics such as FOV and resolution. The layout of the Unity game engine and the experimental setup for this study are depicted in Figure 3.

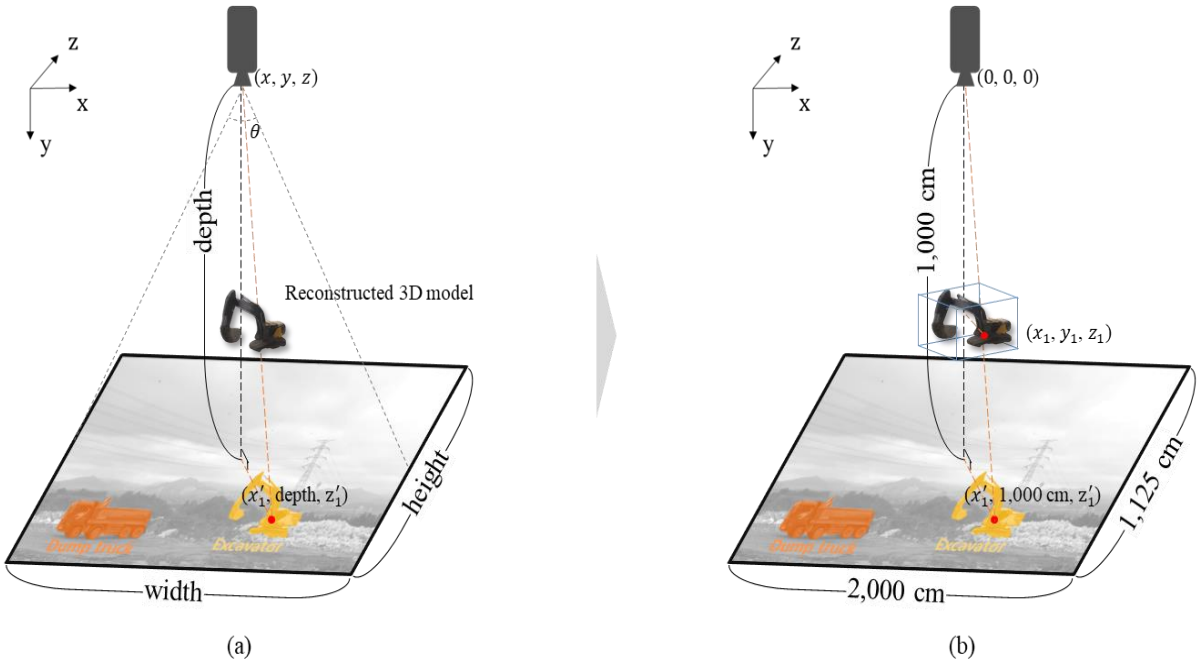

Figure 3. (a) layout of the Unity; (b) experimental setting for this study

Figure 3(a) illustrates the Unity layout, including the camera, objects, and background settings, where θ represents the camera's FOV. The parameters of the virtual camera are adjusted along with the distance between the camera and the background to ensure the camera's output matches the actual recorded footage.

Figure 4 shows the relationship between the camera's FOV in Unity, the distance to the background, and the size of the background. The relationship can be represented as shown in Equation 1. In this study, the camera's FOV was set to 90°, which accordingly sets the depth at 1,000cm from the camera. While the depth can be arbitrarily adjusted according to the user's objectives, the size of the background can be determined based on the resolution of the field image. Given that the resolution of the utilized site video is 1280 × 720, the background's width is calculated to be 2,000cm according to Equation 1. To maintain resolution, the background's height was set at 1,125cm. This depiction is shown in Figure 3(b), where x_1' and z_1', and the size of the respective object, can be known from the bounding box information extracted during the segmentation process. Furthermore, as the size of the 3D reconstructed model can also be obtained in Unity, this information is utilized to calculate the depth value of the 3D reconstructed model (i.e., y_1).

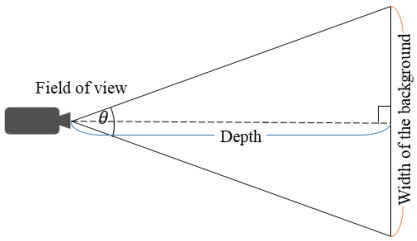

Figure 4. Relationship between the field of view, width of the background, and depth

$$\tan\frac{\theta}{2} = \frac{width\ of\ the\ background/2}{depth} \quad (1)$$

Based on the acquired depth value, x_1 and z_1 are determined through Equation 2.

$$size\ of\ the\ 3D\ model : size\ of\ the\ segmented\ object = y_1 : 1{,}000cm \quad (2)$$

3.3. Work status analysis

At this stage, the objective is to perform status analysis using the 3D information of each object situated in the 3D space. The operational status of the equipment is identified based on several factors: the distance between each object ($\boldsymbol{d}$), the change in the centroid of an object ($\boldsymbol{c}$), and the variation in the angle of the 3D bounding boxes between frames ($\boldsymbol{v}$). In contrast to excavators, which exhibit body rotation, rotational considerations were not factored into the analysis of dump trucks, as they lack distinct rotational operations. The operational status for both excavators and dump trucks are classified into three categories: stationary, moving, and loading. If the distance $\boldsymbol{d}$ between objects falls below a predefined threshold $\boldsymbol{\alpha}$, it is determined that there is interaction between the two objects; otherwise, it is deemed that there is no interaction. For instance, in earthwork operations, the presence of both excavators and dump trucks is essential for progress, meaning that if no interaction is detected, it would be impossible for the operational status of the excavator and dump truck to be classified as loading. The detailed criteria for determining the equipment's operational status under various conditions are presented in Tables 1 and 2.

Table 1. Criteria for analyzing operational status in the presence of interaction

Excavator			Dump truck	
Centroid change ($\boldsymbol{c}$)	Angle variation ($\boldsymbol{v}$)	Operational status	Centroid change ($\boldsymbol{c}$)	Operational status
$\boldsymbol{c} > threshold\ \boldsymbol{\beta}$	-	Moving	$\boldsymbol{c} > \boldsymbol{\beta}$	Moving
		Moving	$\boldsymbol{c} \leq \boldsymbol{\beta}$	Stationary
$\boldsymbol{c} \leq \boldsymbol{\beta}$	$\boldsymbol{v} > threshold\ \boldsymbol{\gamma}$	Loading	-	Loading
	$\boldsymbol{v} \leq threshold\ \boldsymbol{\gamma}$	Stationary	$\boldsymbol{c} > \boldsymbol{\beta}$	Moving
			$\boldsymbol{c} \leq \boldsymbol{\beta}$	Stationary

Table 2. Criteria for analyzing operational status in the absence of interaction

Excavator			Dump truck	
Centroid change ($\boldsymbol{c}$)	Angle variation ($\boldsymbol{v}$)	Operational status	Centroid change ($\boldsymbol{c}$)	Operational status
$\boldsymbol{c} > threshold\ \boldsymbol{\beta}$	-	Moving	$\boldsymbol{c} > \boldsymbol{\beta}$	Moving
		Moving	$\boldsymbol{c} \leq \boldsymbol{\beta}$	Stationary
$\boldsymbol{c} \leq \boldsymbol{\beta}$	$\boldsymbol{v} > threshold\ \boldsymbol{\gamma}$	Moving	$\boldsymbol{c} > \boldsymbol{\beta}$	Moving
			$\boldsymbol{c} \leq \boldsymbol{\beta}$	Stationary
	$\boldsymbol{v} \leq threshold\ \boldsymbol{\gamma}$	Stationary	$\boldsymbol{c} > \boldsymbol{\beta}$	Moving
			$\boldsymbol{c} \leq \boldsymbol{\beta}$	Stationary

4. RESULTS AND DISCUSSION

In this study, experimental data consisted of a one-minute construction site excavation video with a resolution of 1280 x 720 at 30 frames per second. Using the Sam-Track model, object segmentation was performed on 'excavators' and 'dump trucks' within the video, resulting in 1,800 cropped images categorized by equipment type. Examples of the acquired cropped images and the 3D models generated using them are shown in Figure 5.

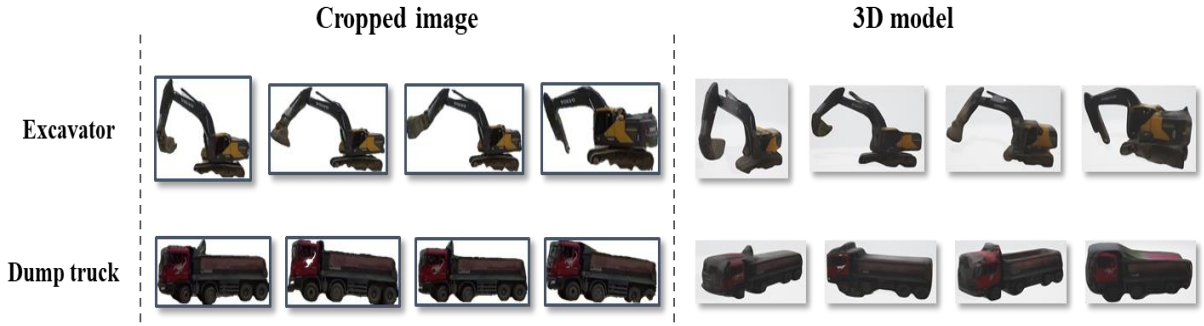

Figure 5. Example of the cropped images and the 3D models

As the results indicate, 3D models could generally be well created using only a single image. However, in the case of excavators, there were instances where certain parts (e.g., the bucket and arm) were missing in the generated models. This issue was identified not as a flaw in the 3D reconstruction model's performance but as a consequence of losses incurred during the segmentation process. To address this, improving the performance of the segmentation or employing 3D shape completion techniques to supplement the missing parts based on a completely generated mesh without losses may be necessary.

After generating the virtual models, they were placed together within Unity's single 3D virtual environment. Subsequently, the 3D information possessed by the objects in the 3D space (e.g., location, object's 3D bounding box) was utilized to analyze the operational status of the objects. In this study, to distinguish the operational statuses outlined in Tables 1 and 2, the threshold value $\boldsymbol{\alpha}$ for determining the distance between objects was set to the length of a dump truck, the threshold $\boldsymbol{\beta}$ for changes in the centroid was 100 cm/sec, and the rotational angle $\boldsymbol{\gamma}$ was set at 15 degrees/sec. Based on these criteria, the work status analysis was conducted at intervals of 0.5 seconds (15 frames), with adjustments made using the analysis values from one second before to one second after each interval to conduct post-correction.

The results demonstrated that the accuracy was 93.33% for excavators and 98.33% for dump trucks, with examples of these outcomes illustrated in Figure 6. The proposed method utilizes the distance between each object ($\boldsymbol{d}$), the change in the centroid of an object ($\boldsymbol{c}$), and the variation in the angle of the 3D bounding boxes between frames ($\boldsymbol{v}$) for the status analysis. Based on these values, the working status of equipment was analyzed within a 3D virtual space. Consequently, even when observing the same object from different camera views and angles, as illustrated in views 1 and 2 of Figure 6, the analysis results will remain consistently uniform. Unlike traditional 2D pixel-based status analysis methods,

there is no need to repeatedly adjust the threshold values of the ruleset based on camera placement or FOV. Additionally, the performance of the developed analysis model is expected to remain relatively high across different application sites.

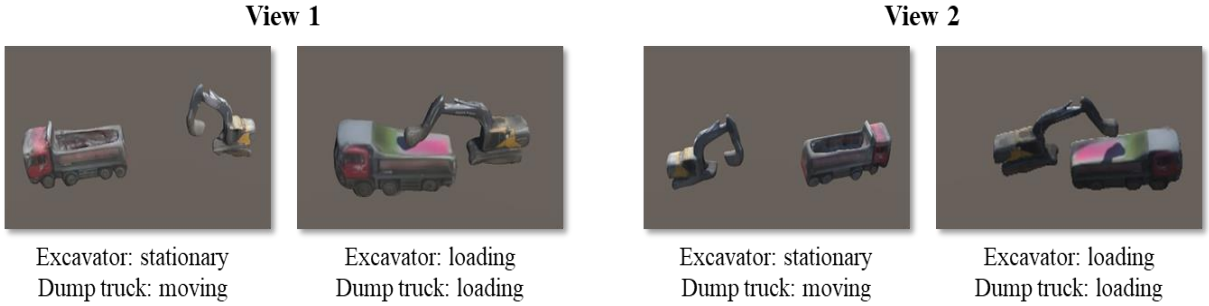

Figure 6. Example of work status analysis results

5. CONCLUSION

In this study, the authors proposed a method for 3D visualization and status analysis of construction site objects using 3D reconstruction technology. The experimental results showed feasibility and satisfactory performance for the proposed method. This study makes the following contributions. First, the proposed method is capable of extracting 3D information of equipment from 2D images using a single camera. Consequently, there is no need for specialized cameras like multi-camera setups, depth cameras, or location sensors. Second, because state analysis is conducted using 3D spatial information, it obviates the need for the manual assumptions required by traditional 2D-pixel information-based analysis. For instance, there is no necessity to search for threshold values of rulesets to optimize performance based on camera placement. In other words, adjustments to the ruleset are not required when applied to different camera views. Finally, the findings of this study can provide meaningful insights and new directions for research in construction video-based 3D information analysis. For example, utilizing this method enables the acquisition of 3D information for each object and its representation within a single virtual space. By representing objects captured from multiple cameras in one virtual environment, this approach has the potential to address issues such as ID switch errors and occlusion problems.

Nevertheless, this study still needs to be improved. In this study, the evaluation was limited to a brief period and focused on a singular construction site. Future research endeavors must ascertain the applicability of the established criteria in ensuring robust performance across extended video sequences and datasets collected from diverse construction environments. A performance comparison with existing 2D pixel-based analysis techniques and addressing the creation of missing 3D models due to poor segmentation results are also necessary.

ACKNOWLEGEMENTS

This work was supported by the National R&D Project for Smart Construction Technology (23SMIP-A158708-04) funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure, and Transport. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00241758 and No. 2021R1A2C2003696).

REFERENCES

[1] A. Panas, J. P. Pantouvakis, “Evaluating research methodology in construction productivity studies”, The Built & Human Environment Review, vol. 3, no. 1, pp. 63-85, 2010.

[2] K. M. El-Gohary, R. F. Aziz, H. A. Adbel-Khalek, “Engineering approach using ANN to improve and predict construction labor productivity under different influences”, Journal of Construction Engineering and Management, vol. 143, no. 8, 04017045, 2017.

[3] E. Rezazaadeh Azar, B. McCabe, “Automated visual recognition of dump trucks in construction videos”, Journal of Computing in Civil Engineering, vol. 26, no. 6, pp. 769-781, 2012.

[4] M. Memarzadeh, M. Golparvar-Fard, J. C. Niebles, “Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors”, Automation in Construction, vol. 32, pp. 24-37,, 2013.

[5] W. Ji, L. Tang, D. Li, W. Yang, Q. Liao, “Video-based construction vehicles detection and its application in intelligent monitoring system”, CAAI Transactions on Intelligence Technology, vol. 1, no. 2, pp. 162-172, 2016.

[6] B. Xiao, S.C. Kang, “Vision-based method integrating deep learning detection for tracking multiple construction machines”, Journal of Computing in Civil Engineering, vol. 35, no. 2, 04020071, 2021.

[7] J. Gong, C. H. Caldas, “Computer vision-based video interpretation model for automated productivity analysis of construction operations”, Journal of Computing in Civil Engineering, vol. 24, no. 3, pp. 252-263, 2010.

[8] H. Kim, S. Bang, H. Jeong, Y. Ham, H. Kim, “Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation”, Automation in Construction, vol. 92, pp. 188-198, 2018.

[9] K. M. Rashid, J. Louis, “Times-series data augmentation and deep learning for construction equipment activity recognition”, Advanced Engineering Informatics, vol. 42, 100944, 2019.

[10] D. Roberts, M. Golparvar-Fard, “End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level”, Automation in Construction, vol. 105, 102811, 2019

[11] J. Kim, S. Chi, J. Seo, “Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks”, Automation in Construction, vol. 87, pp. 297-308, 2018.

[12] M.W. Park, C. Koch, I. Brilakis, “Correlating multiple 2D vision trackers for 3D object tracking on construction sites”, Proceedings of Advancing Project Management for the 21st Century, pp. 560-567, 2010.

[13] 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”, Journal of Computing in Civil Engineering, vol. 32, no. 3, 04018012, 2018.

[14] A. Assadzadeh, M. Arashpour, I. Brilakis, T. Ngo, E. Konstantinou, “Vision-based excavator pose estimation using synthetically generated datasets with domain randomization”, Automation in Construction, vol. 134, 104089, 2022.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.Y. Lo, P. Dollar, “Segment Anything”, arXiv preprint:2304.02643, 2023.

[16] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, Y. Yang, “Segment and track anything”, arXiv preprint:2305.06558, 2023.

[17] M. Liu, R. Shi, L. Chen, Z. Zhang, C. Xu, X. Wei, H. Chen, C. Zeng, J. Gu, H. Su, “One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion”, arXiv preprint: 2311.07885, 2023.