Tracking algorithm for hazardous area in construction sites using PTZ cameras

Jaehwan Seong[1], Hyung-soo Kim[1] and Hyung-Jo Jung[1]

[1]Department of Civil and Environmental Engineering, KAIST, Korea
kevin8633@kaist.ac.kr, rlagudtn0934@kaist.ac.kr, hjung@kaist.ac.kr

Abstract

This paper proposes a PTZ (Pan-Tilt-Zoom) camera-based solution to enhance construction site safety, particularly for small to medium scale projects and those in the early or late stages of development. Traditional sensor-based methods often face challenges related to cost, installation time, and adaptability, limiting their suitability for dynamic construction environments. In contrast, the proposed system leverages PTZ cameras and deep learning techniques specifically Visual Place Recognition (VPR) and Monocular Depth Estimation (MDE) to dynamically redefine hazardous areas in real time as the camera's view changes. The system integrates multiple tracking strategies, including feature-based tracking, scene re-identification, depth-driven adjustments, moving object detection, and image stabilization, to adapt to diverse site conditions. Experimental validation in both virtual and real-world settings demonstrate high accuracy, achieving an average IoU of 0.8627 in virtual sites and 0.7463 in real sites, despite the increased variability in actual conditions. The virtual environment, developed using a game engine, enabled precise simulations of site dynamics such as lighting, weather changes, and object movements. Real site tests further validated the system's practical applicability, highlighting its potential for broader deployment in construction safety management. By enhancing hazard identification and monitoring capabilities, this approach contributes to more effective and comprehensive safety management in construction projects.

Keywords

Safety Management; Computer vision; Real-time hazardous area tracking; Pan-Tilt-Zoom (PTZ) cameras

1 Introduction

The construction industry has long been recognized as a sector with a high risk of occupational accidents due to several intrinsic factors: the use of large-scale equipment, complex processes, collaboration among various stakeholders, and prolonged project durations often conducted outdoors. These characteristics have continually drawn global attention to construction site safety, prompting sustained efforts toward improvement. According to the U.S. Occupational Safety and Health Administration (OSHA), there were 1,075 fatalities in the U.S. construction sector in 2023, accounting for over 20% of all workplace deaths, with falls representing the most frequent cause [1]. In China, data from the Ministry of Housing and Urban-Rural Development (MOHURD) show that between 2012 and 2019, approximately 4,100 accidents claimed 5,011 lives in the construction sector, with falls constituting 52.83% of these incidents. These statistics suggest that accidents frequently occur around static hazard areas (e.g., foundation excavations, edges, openings), underscoring a persistent risk in such environments [2]. Similarly, in Korea, there were 26,829 injuries reported in 2023, accounting for over 20% of all occupational accidents in the construction industry. Small-scale sites often lack systematic safety management measures, leading to a high incidence of falls and collapses [3].

A noteworthy aspect is that many of these fatal incidents occur in relatively predictable "static hazardous areas." That is, certain spatial features such as openings or edges of structures pose an inherent and constant risk throughout the construction process. When this latent hazard is compounded by managerial or technological limitations, the likelihood of accidents increases [4, 5]. As a result, continuously and efficiently monitoring static hazardous areas and detecting unauthorized entries into these areas in real time has emerged as a key priority for global construction safety management.

Against this backdrop, recent research in construction safety management has actively explored the use of computer vision (CV)-based technologies. Traditional manual surveillance in construction sites is labor-intensive, time-consuming, and susceptible to human judgment errors and observer fatigue. By contrast, computer vision technologies employ automated algorithms to analyze targets such as workers, equipment, and structural conditions and have demonstrated utility in various applications, including worker safety monitoring, detecting facility defects, ergonomic posture assessments, and process monitoring [6-10]. Recently, approaches aimed at the real-time recognition of and intrusion detection into static hazardous areas have also emerged, such as the work by Mei et al., which proposed a more clearly defined hazardous area identification model based on computer vision [11]. Their system uses object detection to locate workers in each video frame and then determines the level of risk by checking whether the detected position falls within a polygon that represents a predefined hazardous area on the camera image.

However, most existing detection techniques for static hazardous areas rely on images from fixed cameras, requiring manual designation of these areas. The manually designated area is defined based on the image coordinates of the camera, and when the camera moves, the dangerous area is distorted and the worker's entry into the wrong area is detected. This approach presents significant constraints when applied in real-world scenarios. Nowadays construction sites are vast, and observation points change frequently as the project progresses, making it challenging to comprehensively monitor all areas through a static viewpoint. While the use of Pan-Tilt-Zoom (PTZ) cameras with adjustable fields of view is on the rise, each camera movement necessitates redefinition of the static hazardous area, a process that is both inefficient and lacking in adaptability as shown in Figure 1. In other words, current methods suffer from limited dynamic adaptability to changes in camera orientation, undermining both the reliability and practicality of real-time hazard monitoring.

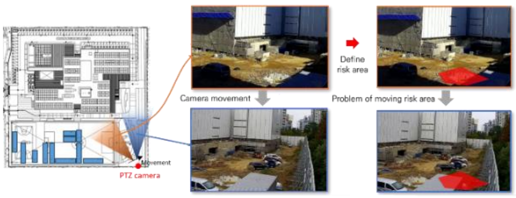

Figure 1. Problem of moving to a hazardous area

To address these limitations, this study proposes a safety management framework that estimates the pose of a PTZ camera using image data alone and automatically reconfigures predefined static hazardous areas according to camera movements. The key contribution lies in extracting spatial information solely from image input, without relying on additional sensors (e.g., IMUs), and using this information to dynamically map static hazardous areas in response to camera adjustments. As a result, the system can continuously track previously defined static hazard areas and determine whether intrusions occur, regardless of the camera's position and viewing angle. This approach is anticipated to improve both the operational flexibility of camera usage and the accuracy of monitoring in practical environments, thereby advancing real-time safety management systems that focus on static hazards. By applying the system proposed by Mei et al. to the tracked risk area, workers' access to the hazardous area can be detected. Moreover, by expanding the applicability of CV-based construction safety monitoring, this research may ultimately provide a practical method for managing various forms of risks across diverse project scales and processes in the future.

2 Proposed Framework

Construction sites are operated over extended periods, during which camera imagery is subject to significant variability due to equipment and worker movement, changes in workflows, illumination fluctuations, and weather conditions. To reliably track predefined hazardous areas under such variability, it is necessary to utilize robust algorithms against various factors such as camera motion, lighting changes, object rearrangements, and weather fluctuations. In this study, we propose a framework that, using only the footage captured from a PTZ (Pan-Tilt-Zoom) camera, estimates the camera pose and automatically reconfigures a predefined static hazardous area according to changes in the camera's viewpoint (see Figure 2).

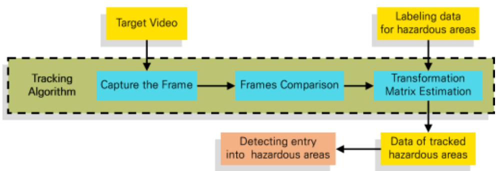

Figure 2. Framework of tracking algorithm for hazardous area

The proposed framework consists of three key stages:

1. **Camera poses estimation:** Using feature points extracting and matching technique, a transformation matrix between two viewpoints is computed. This method ensures real-time capability while maintaining relatively stable features matching under various environmental changes.
2. **Site verification based on Visual Place Recognition (VPR):** A VPR network is employed to determine whether the current image is from the same site where the hazardous area was previously defined. This allows for temporarily halting hazardous area tracking whenever the camera angle or position changes causes a completely different location to be captured, preventing safety incidents due to wrong identification.
3. **Robustness enhancement using depth information (MDE):** Depth information estimated

from a single image by a Monocular Depth Estimation (MDE) network is integrated into the tracking process. This approach enables stable background feature extraction despite lighting changes or temporary occlusions, thereby offering more robust adaptation to environmental changes when tracking the actual hazardous area.

2.1 Overview of the Framework

Figure 2 shows the overall structure of the proposed framework. First, features are extracted from images at two different points in time, then matched to estimate a transformation matrix. Traditional methods such as SIFT, SURF, and ORB are relatively easy to implement but are vulnerable to lighting changes and often struggle with speed-accuracy trade-offs. In contrast, the lightweight deep learning-based approach (XFeats) proposed by Potje et al. enables a better balance between real-time processing and accuracy [12]. By estimating camera pose changes between frames using this method, the static hazardous area is reprojected onto the image according to camera movements, enabling continuous tracking.

However, relying solely on feature matching can lead to erroneous matches caused by environmental changes (fog, rain, occlusion, changes in wall layout), ultimately resulting in incorrect hazardous area projections as shown in Figure 3. To mitigate this, the VPR network is employed to determine whether the captured footage is from the same site. If it is not, hazardous area tracking is suspended to prevent safety issues stemming from erroneous tracking. In addition, MDE is utilized to strengthen the structural understanding of the background beyond mere image-based features. By integrating depth information obtained from a single image, the framework achieves greater robustness against environmental changes during feature matching and VPR.

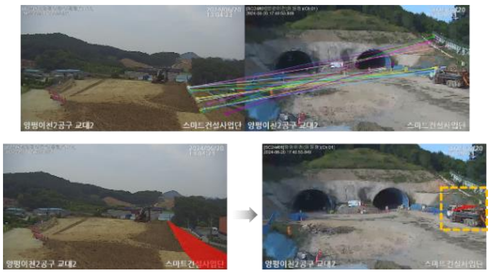

Figure 3. Failure of tracking due to PTZ camera operation (red area)

2.2 Estimation of Relative Camera Pose

Camera poses estimation is based on feature extraction and matching. Traditional methods (e.g., SIFT, SURF, ORB) are human-interpretable and easy to implement, but are sensitive to lighting and scale changes, and present a trade-off between speed and accuracy. Following Melekhov et al. CNN-based feature extraction and matching techniques have been explored [13]. Potje et al. introduced a lightweight approach, XFeats, combining real-time performance with accuracy [12]. In this study, XFeats is adopted to ensure stable tracking under diverse environmental changes in large-scale construction sites. Using the estimated transformation matrix, the camera pose is continuously updated, and the static hazardous area is reconfigured in real-time.

2.3 Visual Place Recognition (VPR)

VPR is a core component for identifying whether the captured images are from the same site where the hazardous area was previously defined. Traditional feature-based methods are highly vulnerable to changes in lighting (day/night) and weather (clear/rain/snow), resulting in misinterpretations that consider the same site as a completely different space. To address this issue, this study references the latest CNN- and Transformer-based VPR approaches [14, 15]. By employing the MixVPR method, high-dimensional features are mixed to generate a global descriptor. The similarity between global descriptors allows robust determination of whether the current site matches the previously defined hazardous area location.

Using the VPR results, the proposed framework maintains hazardous area tracking if it is the same site or suspends tracking or updates the reference images if not, preventing safety management issues caused by erroneous matches. The incorporation of MDE-derived depth information assists in robustly capturing background structures, further enhancing resilience to weather changes or temporary occlusions and improving overall recognition performance.

2.4 Monocular Depth Estimation (MDE)

Monocular Depth Estimation (MDE) is a technique that estimates the distance to each pixel in an image using only a single RGB camera input. It can be leveraged to acquire three-dimensional environmental information from construction sites without expensive equipment or complex installations, effectively serving as a substitute for RGB-D cameras. By employing an encoder-decoder neural network architecture, spatial features are extracted, and depth is predicted from the input image. Incorporating LSTM can capture temporal features, enabling more stable depth estimation. In addition, training on ground-truth depth maps that incorporate physical constraints ensures that the estimated depth is robust to environmental changes such as lighting variations, weather conditions, and occlusions.

In this study, a Monocular Depth Estimation (MDE) network with a DINOv2 encoder and a DPT decoder is employed. We use a model pretrained on large-scale unlabeled data as demonstrated by Yang et al. [16]. This provides effective depth estimation, which, when integrated into feature matching and VPR, enhances the stability and accuracy of hazardous area tracking in construction sites.

3 Experiments

This chapter explains the preparations and approaches used in the experiments for the proposed PTZ camera-based static hazardous area tracking algorithm. It also defines the evaluation indicators that will be used to interpret the experimental results presented in Section 4. To overcome the challenges of collecting and labelling video data from actual construction sites while minimizing safety management burdens, a virtual construction site environment was developed using the Unity game engine. This approach enabled the simulation of realistic variables, including camera operations (pan, tilt, zoom), weather, and lighting changes, while ensuring accurate labelling information and controlled hazardous area configurations. The final performance was verified by applying the approach, which showed the highest performance in the virtual construction site, to the actual field.

3.1 Implementation Details

Section 3.1 describes the implementation details of the algorithms used in the experiment, the configuration of the virtual environment, and the process of labeling hazardous areas.

3.1.1 Virtual Construction Site Implementation

A virtual construction site was built using Unity as shown as Figure 4, incorporating various architectural elements (structures, equipment, terrain features, etc.), changes in lighting over time (day/night transitions), weather variations (rain, fog), and object movements. A PTZ (Pan-Tilt-Zoom) camera was set up to mimic actual site operations by randomly altering camera poses. Objects, such as equipment and workers, moved unpredictably, reflecting the dynamic nature of active construction sites.

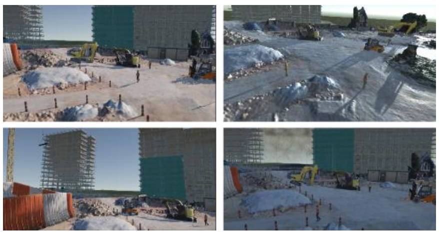

Figure 4. Virtual construction site

Camera operation times along a virtual timeline were categorized into four segments: less than 10 seconds (brief), 10 to 30 seconds (short), 30 seconds to 1 minute (mid), and over 1 minute (long). This categorization aimed to evaluate tracking performance over extended periods.

3.1.2 Hazardous Area Labeling and Ground Truth Generation

Within the virtual environment, certain areas were predefined as “static hazardous areas” and used as ground truth. As the camera moved, the projective representation of these hazardous areas changed. The proposed algorithm continuously reprojects these areas in real-time based on camera movement. The lighting environment was shifted over a 100 s day/night cycle, and intermittent rainfall was introduced to replicate weather-induced visibility changes. The ground truth information was utilized in subsequent evaluations (see Section 3.3 for evaluation metrics).

3.1.3 Detail Parameter of Applied Model

Feature extraction and matching for localization and tracking are performed using XFeats, a CNN-based feature extraction and matching model designed for real-time operation in resource-constrained environments. XFeats employs a lightweight CNN backbone that maintains a low number of channels, optimizing computational efficiency while preserving high-resolution feature representations. The model was pretrained on a combination of the Megadepth dataset and synthetically warped images from COCO to enhance its robustness. To further improve structural feature representation, MixVPR, a CNN-based descriptor specifically designed to handle severe illumination and weather variations, is utilized. MixVPR was pretrained on the Pitts30k-train dataset, and its encoder is employed to extract descriptors that capture the structural characteristics of each image. The cosine similarity between these descriptors is then used to assess structural similarity across images, enabling the determination of on/off states for risk area tracking. Additionally, for monocular depth estimation (MDE), a pretrained model was adopted following the approach proposed by Yang et al., as discussed in Section 2.4.

3.2 Experimental Setup and Approaches

This section outlines various approaches used to implement the proposed tracking mechanism. These include frame-to-frame comparison techniques, the use of a reference image, VPR (Visual Place Recognition) methods, and strategies for flexible hazardous area activation/deactivation by incorporating MDE-based depth information. Five approaches were proposed, and the structure of each is illustrated in Figure 5.

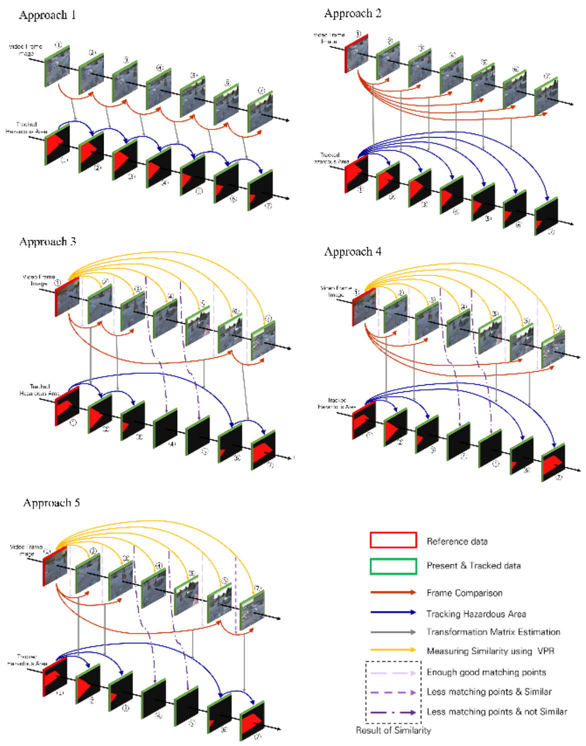

Figure 5. Architecture of each approach

3.2.1 Simple Transformation Approaches (Approaches 1 & 2)

Approach 1 (Continuous Frame Comparison) involves estimating the transformation matrix between consecutive frames for hazardous area tracking. This approach, commonly used in SLAM, effectively handles small incremental changes but suffers from cumulative errors over time, presenting challenges for long-term tracking.

Approach 2 (Reference Image-Based) defines the first frame as a reference and directly estimates the transformation matrix between the current frame and the reference frame for reprojection. This approach avoids long-term error accumulation but may face difficulties when the scene differs significantly from the reference image, making direct matching challenging.

3.2.2 VPR for Scene Change Adaptation (Approaches 3 & 4)

In construction sites, the camera may capture different scenes where the previously defined hazardous area no longer applies. To address this, 2 types of VPR methods were applied. Type 1 determines similarity based on feature points, while Type 2 integrates image descriptors with depth information from MDE. By evaluating whether the current frame and the reference frame depict the same location, the hazardous area can be deactivated when shooting a different site, mitigating tracking errors caused by scene changes. Approach 3 is a method of introducing the VPR plan based on Approach 1, and Approach 4 is a method of introducing the VPR plan based on Approach 2

3.2.3 Hybrid Approach (Approach 5)

Approach 5 overcomes the limitations of Approaches 1–4 by dynamically selecting between reference image-based tracking and VPR-guided results. If enough matching points are extracted, reference image-based tracking (Approach 2) is used; otherwise, previous-frame-based tracking (Approach 1) is employed. At the same time, VPR results guide the deactivation of hazardous areas when the scene changes. The integration of VPR and MDE enables stable activation and deactivation of hazardous areas, even under varying environmental conditions.

3.3 Evaluation Indicators

In this section, we define the quantitative evaluation metrics for interpreting the experimental results presented in Section 4.

3.3.1 Intersection over Union (IoU)

The accuracy and stability of hazardous area tracking are evaluated using various indicators. The performance of hazardous area tracking is quantitatively assessed using the Intersection over Union (IoU) metric and several supplementary indicators. First, the IoU (see Equation (1)) is the most representative indicator of similarity between the actual hazardous area (R_t) and the tracked hazardous area (R_g). When comparing the tracking results and the actual hazardous area in each frame.

$$IoU = \frac{R_t \cap R_g}{R_t \cup R_g} \quad (1)$$

An IoU value closer to 1 indicates a higher overlap between the two areas, signifying a closer match between the tracked and actual hazardous areas. However, to comprehensively evaluate performance across the entire video, a single frame value of IoU is insufficient. Therefore, the following additional indicators are employed to assess overall tracking performance and reliability. This metric evaluates the similarity between the ground truth hazardous area and the area reprojected by the tracking algorithm. It represents the ratio of the intersection to the union of the two areas, and a higher IoU value indicates greater tracking accuracy. Through the average IoU across all frames ($\overline{IoU}$), the ratio of frames where $IoU \geq 0.80$ (IoU_{80}), and the ratio of frames where $IoU \geq 0.95$ (IoU_{95}) the tracking performance can be evaluated from multiple perspectives.

3.3.2 Accuracy of Activation and Deactivation

When applying the VPR algorithm, changes in the scene can require the hazardous area to be deactivated; AoD measures the accuracy of deactivation in such cases. Conversely, in situations where the hazardous area should remain active, AoA indicates how accurately it remains activated. These metrics allow for an assessment of scene change adaptability and overall reliability in real-world site management.

3.3.3 Long-term Operational Stability

The degradation of the $\overline{IoU}$ based on the length of the video (brief, short, mid, and long) provides a view of the long-term tracking stability. We imply that the approach that does not significantly degrade the $\overline{IoU}$ for long videos compared to brief videos is the approach to maintain consistent performance in long-time operations. Therefore, the approach that does not significantly degrade for long videos compared to brief videos is more advantageous for construction site applications that require real-world long-term operations. These complementary metrics are used to compare and evaluate the performance of the proposed tracking Approaches (Approaches 1-5). Specific performance figures and comparison results are discussed in detail in Section 4.

4 Results

4.1 Results at Virtual Construction Site

4.1.1 Comparison of Basic Tracking Methods

Based on the $\overline{IoU}$, Approach 2, which tracks using a reference frame, showed a slightly better performance than Approach 1, which compares consecutive frames. In short and mid-length video scenarios, Approach 2 outperformed Approach 1. However, in long videos, both approaches exhibited similar performance, while in brief videos, Approach 1 performed slightly better. Overall, Approach 2 holds an advantage in terms of $\overline{IoU}$, but each approach comes with its own strengths and weaknesses when examining additional indicators. The performance of Approach 1 and Approach 2 is shown in Table 1 and shown as Figure 6.

4.1.2 Effect of Introducing VPR

Approaches 3 and 4 improved the $\overline{IoU}$, IoU_{80} and IoU_{95} compared to their respective counterparts, Approaches 1 and 2. Notably, Approach 4 demonstrated superior overall performance compared to Approach 3. However, in both approaches, AoA was observed to be significantly degraded compared to Approach1.

When comparing VPR types (Types 1 and 2), Type 1 accurately identified deactivation timing but often excessively deactivated hazardous areas, reducing its AoD value. In contrast, Type 2 achieved a higher AoA value, allowing for more reliable detection of actual worker access. However, since neither directly accounts for transformation inaccuracies, there are some limitations in determining the precise deactivation timing. The performance of each approach is detailed in Tables 2 and 3 and shown as Figure 6.

Table 1. Performance by tracking Approach 1 and 2

Tracking Method		Approach 1	Approach 2
$\overline{IoU}$		0.7214	0.7460
$\overline{IoU}_{type}$	brief	0.8513	0.8165
	short	0.7859	0.9127
	mid	0.7372	0.7908
	long	0.6898	0.6833
IoU_{80}		54.54%	76.11%
IoU_{95}		39.43%	54.92%
AoD		90.13%	58.80%
AoA		97.87%	89.73%

Table 2. Performance by tracking Approach 3

Tracking Method		Type 1	Type 2
$\overline{IoU}$		0.5988	0.6908
$\overline{IoU}_{type}$	brief	0.8193	0.7848
	short	0.6876	0.7823
	mid	0.5895	0.7110
	long	0.5662	0.6541
IoU_{80}		52.92%	53.26%
IoU_{95}		42.48%	38.59%
AoD		96.48%	87.84%
AoA		56.77%	86.82%

Table 3. Performance by tracking Approach 4

Tracking Method		Type 1	Type 2
$\overline{IoU}$		0.8326	0.7606
$\overline{IoU}_{type}$	brief	0.8820	0.8355
	short	0.9143	0.9168
	mid	0.8714	0.7934
	long	0.7930	0.7055
IoU_{80}		84.68%	77.58%
IoU_{95}		65.66%	56.44%
AoD		99.89%	62.98%
AoA		77.62%	89.46%

4.1.3 Performance of the Hybrid Approach

Approach 5 is a hybrid approach that dynamically decides whether to use the reference frame based on the number of matching points and relies on VPR results for activation and deactivation. When combined with descriptor-based VPR, this method achieved the highest $\overline{IoU}$ and most stable performance metrics overall. Approach 5 recorded top scores in most of the evaluated indicators. This highlights that integrating depth information and adopting a flexible reference-frame selection strategy significantly improves long-term stability and accuracy in hazardous area tracking, especially under the complex, dynamic conditions of construction sites. Detailed performance results for each method can be found in Table 4 and shown as Figure 6.

Table 4. Performance by tracking Approach 5

$\overline{IoU}$		0.8627
$\overline{IoU}_{type}$	brief	0.9144
	short	0.9322
	mid	0.8789
	long	0.8360
IoU_{80}		87.12%
IoU_{95}		66.26%
AoD		97.20%
AoA		87.11%

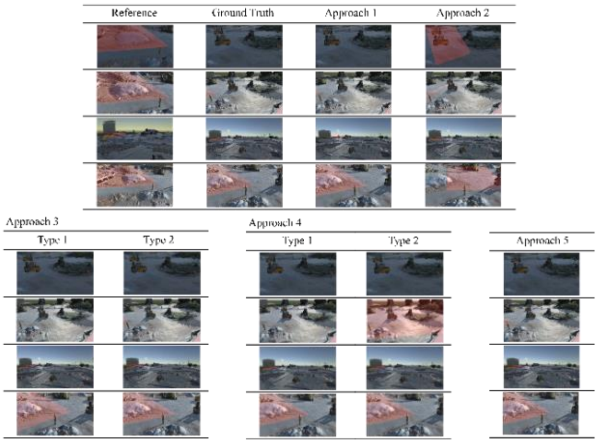

Figure 6. Tracking result at the virtual site

4.2 Results at Real Construction Site

The proposed algorithm (Approach 5 which is the highest performance method in virtual sites, illustrated in Figure 7) was validated using a PTZ camera at an actual construction site. For 20 videos collected from the PTZ camera at 13 sites, the initially set hazardous area and the hazardous area moved after the PTZ camera operation were manually labeled. In 20 videos, the hazard areas were manually defined based on the first frame, and the ground truth for the final tracking results was also manually established using the last frame. To evaluate the tracking performance, we calculated the IoU between the ground truth and the hazard areas tracked by the proposed method.

Performance was measured by comparing the hazardous area tracked by the algorithm with the ground truth of the hazardous area after PTZ camera operation. The tracking algorithm achieved average performance score of 0.7463 across 20 videos, with detailed results illustrated in Figure 8.

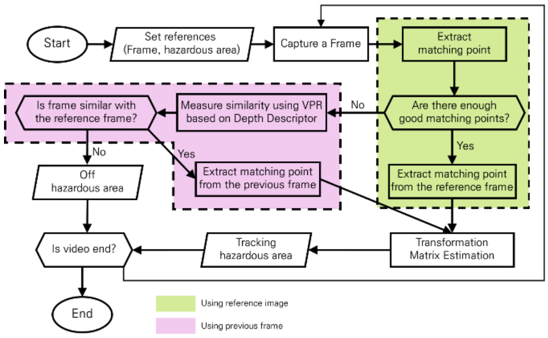

Figure 7. Flow chart of algorithm to track the hazardous area for PTZ camera

The complexity of real sites that was not considered in the virtual sites resulted in a lower result than the performance in the virtual sites. In general, segmentation models evaluated using the IoU metric are considered highly accurate when mean IoU exceeds 0.7. Based on this standard, our proposed algorithm appears sufficiently robust for hazard zone tracking in construction sites. However, from a safety management perspective, further improvements in tracking accuracy will be necessary for more precise hazard detection.

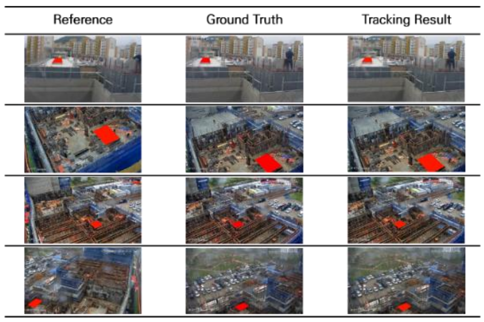

Figure 8. Tracking result at the real site

5 Conclusion

In this study, a hazardous area tracking algorithm was developed for PTZ cameras operated at construction sites. The tracking algorithm's performance was validated in both simulated and real-world construction site environments. The proposed algorithm was implemented by extracting well-matched point pairs from the reference image, deciding on reference image usage for transformation estimation, and activating hazardous areas based on descriptor cosine similarity. The proposed algorithm was verified at a virtual construction site and actual construction.

The proposed system achieved a tracking performance score of 0.8627 and an IoU_{80} of 87.12% in virtual construction site simulations. In addition, by applying the estimation algorithm to actual construction sites, the average IoU was 0.7463, confirming that it could be properly applied to the actual site. As a result, a smart safety system that can detect construction workers' access to hazardous areas can be applied to PTZ cameras, thereby enhancing the safety of construction sites. The system is also expected to enable the projection of virtual object information for construction site safety management without requiring separate control points or sensors, thus contributing to more efficient site management.

This study focused solely on tracking methods for 2D objects. However, hazardous areas at actual construction sites often appear in the form of 3D, and it is necessary to implement the tracking function for 3D objects in subsequent studies. Additionally, while the hazardous area tracking algorithm was implemented, future research should evaluate its accuracy in detecting worker access to hazardous areas at real construction sites, especially when integrating previously proposed access detection algorithms.

Acknowledgements

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2020-KA156208)

References

[1] OSHA (Occupational Safety and Health Administration). Commonly used statistics. On-line: https://www.bls.gov/news.release/cfoi.nr0.htm, Accessed: May 03, 2025.

[2] MOHURD (Ministry of Housing and Urban-Rural Development). MOHURD’s notification on the safety accidents in housing and municipal works in 2019. [In Chinese.] Stand. Eng. Constr. 2020 (7): 51–53, 2020.

[3] Kim, D. Status of industrial accidents in 2021. [In Korean.] Ulsan: Korea Occupational Safety & Health Agency. Status of industrial accidents in 2023, 2024

[4] Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., & Li, C. Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment. Automation in Construction, 93, 148-164, 2018

[5] Fang, W., Zhong, B., Zhao, N., Love, P. E., Luo, H., Xue, J., & Xu, S. A deep learning-based approach for mitigating falls from height with computer vision: Convolutional neural network. Advanced Engineering Informatics, 39, 170-177, 2019

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

[7] Kim, H. [Construction Safety A to Z] Smart Construction Safety Using AI. [In Korean.] Daehan Economic Daily, 2023

[8] Koch, C., Georgieva, K., Kasireddy, V., Akinci, B., & Fieguth, P. A review on computer vision-based defect detection and condition assessment of concrete and asphalt civil infrastructure. Advanced engineering informatics, 29(2), 196-210, 2015

[9] Ray, S. J., & Teizer, J. Real-time construction worker posture analysis for ergonomics training. Advanced Engineering Informatics, 26(2), 439-455, 2012

[10] Han, K. K., & Golparvar-Fard, M. Appearance-based material classification for monitoring of operation-level construction progress using 4D BIM and site photologs. Automation in construction, 53, 44-57, 2015

[11] Mei, X., Zhou, X., Xu, F., & Zhang, Z. Human intrusion detection in static hazardous areas at construction sites: Deep learning–based method. Journal of Construction Engineering and Management, 149(1), 2023

[12] Potje, G., Cadar, F., Araujo, A., Martins, R., & Nascimento, E. R. XFeat: Accelerated Features for Lightweight Image Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024

[13] Melekhov, I., Ylioinas, J., Kannala, J., & Rahtu, E. Relative camera pose estimation using convolutional neural networks. In Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings 18 (pp. 675-687). Springer International Publishing, 2017

[14] Zhu, S., Yang, L., Chen, C., Shah, M., Shen, X., & Wang, H. R2former: Unified retrieval and reranking transformer for place recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023

[15] Ali-Bey, A., Chaib-Draa, B., & Giguere, P. Mixvpr: Feature mixing for visual place recognition. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023

[16] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., & Zhao, H. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024