



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Master of Science

**Single Camera-based 3D Pose Estimation
and Localization of Construction Equipment
using Virtual Models**

February 2021

**Department of Civil & Environmental Engineering
The Graduate School
Seoul National University**

Junghoon Kim

**Single Camera-based 3D Pose
Estimation and Localization of
Construction Equipment using Virtual
Models**

지도교수 지 석 호

이 논문을 공학석사학위논문으로 제출함

2021년 1월

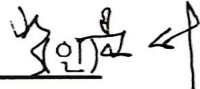
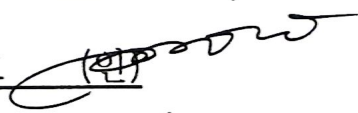

서울대학교 대학원

건설환경공학부

김 정 훈

김정훈의 석사학위논문을 인준함

2021년 1월

위원장	<u>박문서</u> 
부위원장	<u>지석호</u> 
위원	<u>박만우</u> 

Abstract

Single Camera-based 3D Pose Estimation and Localization of Construction Equipment using Virtual Models

Junghoon Kim

Department of Civil & Environmental Engineering

The Graduate School

Seoul National University

In recent years, there is increasing practical interest in construction site monitoring using closed circuit television (CCTV) camera installed on site. Various image analysis technologies have also been developed for image monitoring of construction sites, such as activity classification and productivity analysis of construction equipment using CCTV image data, and detection of access to dangerous areas. However, previous studies mainly focused on identifying and tracking construction equipment or classifying activity in CCTV image data, but did not deal with detailed pose information of construction equipment.

Pose estimation of construction equipment is the acquisition of two-dimensional (2D) or three-dimensional (3D) coordinate information (i.e., its location and orientation) for each keypoint of the equipment, providing basic mechanical information for identifying construction equipment posture, remote control, productivity analysis of construction projects, and safety analysis of construction sites. Knowing construction equipment's pose information, it is possible to understand the operation of the equipment in more detail and detect hazardous conditions.

This research proposes a method to 3D pose estimation and localization of construction equipment using a single camera image data and 3D virtual model. The research process consists of four main steps as follows. First, the construction site monitoring using the existing image analysis technology and construction equipment pose estimation were defined from literature research. Second, 3D virtual model's keypoints were stored in the 2D image data of the construction site through 2D - 3D annotation process. Third, a construction equipment pose estimation method was developed that detects and extracts construction equipment from the construction site image data, derives similar images through image matching, and estimates the equipment pose. Finally, the performance of the construction equipment pose estimation result was evaluated using image data collected at three construction sites.

As a result, the pose of construction equipment can be estimated from a single camera image data through the method proposed in this research, and it was confirmed that pose estimation is possible even for the part where the equipment is obscured. In addition, it is expected to be helpful in analyzing interactions between the equipment by using pose estimation results.

Keywords: Single Camera-based, 3D Virtual Model, 3D Pose Estimation,
Construction Equipment

Student Number: 2019-27248

Contents

Chapter 1. Introduction	1
1.1 Research Background	1
1.2 Problem Statement	3
1.3 Research Objectives and Scope	4
1.4 Dissertation Outline	6
Chapter 2. Theoretical Background and Related Work	
.....	8
2.1 Vision-based Construction Site Monitoring	9
2.2 Pose Estimation in Construction	12
2.2.1 Internet of Things-based Approach	12
2.2.2 Vision-based Approach	15
Chapter 3. Construction Equipment's 3D Pose Estimation	
.....	17
3.1 2D - 3D Annotation	19
3.2 Development of Pose Estimation Method	24
3.2.1 Object Detection	24
3.2.2 Image Matching	29
3.2.3 Pose Estimation and Localization	33

Chapter 4. Experimental Results and Discussions	48
4.1 Object Detection and Image Matching	49
4.2 Pose Estimation and Localization.....	54
4.3 Visualization	60
4.4 Evaluation of Validity	63
4.5 Discussion.....	65
Chapter 5. Conclusions.....	66
5.1 Summary and Contributions.....	66
5.2 Limitation and Future Study	68
Bibliography.....	70
Abstract (Korean).....	78

List of Tables

Table 3.1	Stored pose information example by 2D - 3D annotation (cm)..23
Table 4.1	Performance of trained object detection model (AP)49
Table 4.2	Example of pose estimation in the test dataset (Site 1, excavator)55
Table 4.3	Pose interpolation results from no.1 to p_156
Table 4.4	Pose estimation results from no.40 to no.60 (Site 1, excavator)57
Table 4.5	Limits of pose change during 30 frames in train dataset (cm)...58
Table 4.6	Pose interpolation results from no.20 to no.4359
Table 4.7	RMSE for each keypoint (excavator)64
Table 4.8	RMSE for each keypoint (dump truck).....64

List of Figures

Figure 1.1	Camera layouts (previous approaches vs in real)	3
Figure 1.2	Overview of research methodology	5
Figure 3.1	3D virtual models of excavator and dump truck	19
Figure 3.2	Keypoints of excavator 3D model	20
Figure 3.3	Keypoints of dump truck 3D model	20
Figure 3.4	Unity game engine screen	21
Figure 3.5	2D - 3D annotation examples	22
Figure 3.6	Faster region-based convolutional neural network model	25
Figure 3.7	Examples of train image and test image sets for each site	26
Figure 3.8	Crop image database example (Site 1)	27
Figure 3.9	Crop image database example (Site 2)	28
Figure 3.10	Crop image database example (Site 3)	28
Figure 3.11	Template matching algorithm	29
Figure 3.12	Image matching algorithm	32
Figure 3.13	Example of pose information stored in the image matching result	33
Figure 3.14	Same pose will be taken regardless of crop image's location	34
Figure 3.15	Triangle AA similarity	36
Figure 3.16	Layout in Unity game engine space	37
Figure 3.17	AA similarity condition to calculate (x_1, y_1, z_1)	39
Figure 3.18	AA similarity condition to calculate (x_2, y_2, z_2)	42
Figure 3.19	Flowchart of the entire process of correcting by removing outliers	44
Figure 4.1	Examples of experimental results (Site 1)	49
Figure 4.2	Examples of experimental results (Site 2)	50

Figure 4.3	Examples of experimental results (Site 3)	50
Figure 4.4	Examples of image matching results (excavator).....	52
Figure 4.5	Examples of image matching results (dump truck)	53
Figure 4.6	Process of performing pose estimation and localization	54
Figure 4.7	Example of 2D visualization (Site 1, dump truck)	60
Figure 4.8	Example of 3D visualization (Site 1).....	61
Figure 4.9	Example of 3D visualization (Site 2).....	62
Figure 4.10	Example of 3D visualization (Site 3).....	62

Chapter 1. Introduction

1.1 Research Background

In recent years, with the rapid development of image analysis technology, there is increasing practical interest in construction site monitoring using closed circuit television (CCTV) camera installed on site. Many researchers have also developed various image analysis technologies for construction site monitoring, such as activity recognition, productivity analysis of construction equipment, and detection of access to hazardous areas using CCTV image data (Chen et al. 2020, Kim et al. 2020, Li et al. 2015). However, previous studies have mainly focused on identifying, tracking, and classifying activity of construction equipment at construction sites using image analysis technology. There have been some limited studies to estimate the pose information of construction equipment to understand the dynamic state of the equipment and the posture change of construction equipment.

Pose estimation of construction equipment is the acquisition of two-dimensional (2D) or three-dimensional (3D) coordinate information (i.e., its location and orientation) for each keypoint of the equipment, providing basic mechanical information for identifying construction equipment posture, remote control, productivity analysis of construction projects, and safety analysis of construction sites (Tang et al. 2020). Knowing the pose

information of construction equipment, it is possible to understand the operation of the equipment in more detail and detect hazardous conditions (Vahdatikhaki et al. 2015, Luo et al. 2020). For example, collision accidents are often caused by movement of the body as well as the boom and arm of an excavator. However, the equipment's pose information can explain the dynamic state of the entire equipment based on the movement of equipment's keypoints. Thus it is possible to prevent potential collisions (Luo et al. 2020, Yuan et al. 2017). In addition to monitoring techniques that recognize equipment work, studies have recently been conducted to estimate the 3D pose of equipment, but there are limitations in actual field applications such as the difficulty of setting up numerous cameras at the site to take pictures of one equipment at the same time, or the difficulty of estimating poses when parts of the equipment are obscured (Soltani et al. 2018, Liang et al. 2019, Lundeen et al. 2016).

1.2 Problem Statement

Previous studies on construction site monitoring using image analysis technology mainly focused on identifying, tracking equipment, and classifying work. Therefore, there is a lack of understanding of the dynamic state of the entire equipment. Knowing the pose information of construction equipment provides a more detailed view of the dynamic state and posture changes of the equipment, and it is possible to prevent potential collisions with nearby workers or equipment.

To this end, many researches have been conducted to estimate the 3D pose of construction equipment using internet of things (IoT) sensors or vision, but there are limitations to actual field application such as the need to install sensors for all driving parts of the equipment or the need for numerous camera layouts that do not match the real site situation (Figure 1.1).

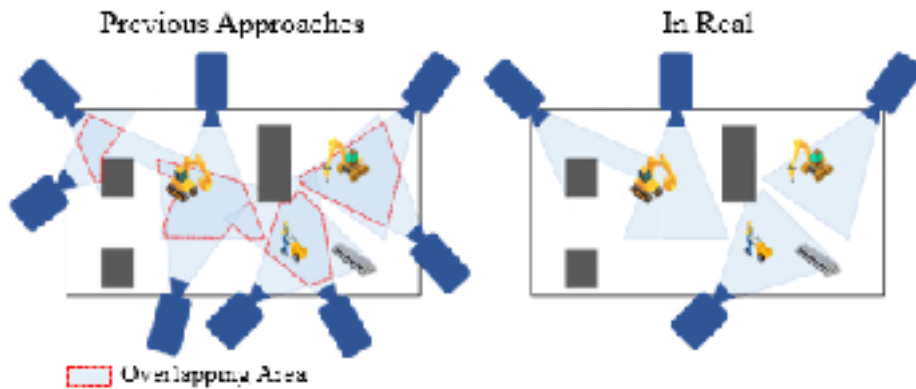


Figure 1.1 Camera layouts (previous approaches vs in real).

1.3 Research Objectives and Scope

This research aims to estimate and localize the 3D pose of construction equipment from a single camera image using virtual models. By expressing the equipment in 3D virtual space, it is possible to check the situation of the entire site, the movement of the equipment, and the interaction with each other. And among the construction equipment used in the construction project, excavator and dump truck were selected as analysis targets because they have a significant impact on productivity and safety of construction project, and interact with each other.

A research methodology (Figure 1.2) and the specific objectives to achieve the primary objective are as follows:

1) Objective 1: Set the keypoints required to understand the dynamic state of the entire equipment and to identify posture changes. And annotate 2D image (construction site image data) with 3D virtual model of construction equipment.

2) Objective 2: Train the object detection model to detect analysis targets (i.e., excavator, dump truck) from construction site image data. Next, extract analysis targets and build a crop image database (DB) for image matching.

3) Objective 3: Detect and extract analysis targets from the construction site image data using the trained object detection model, and derive the most similar image from the crop image database through image matching.

4) Objective 4: Estimation and localization of 3D pose of analysis targets in construction site image data using 3D pose information for virtual models annotated to image matching derived result.

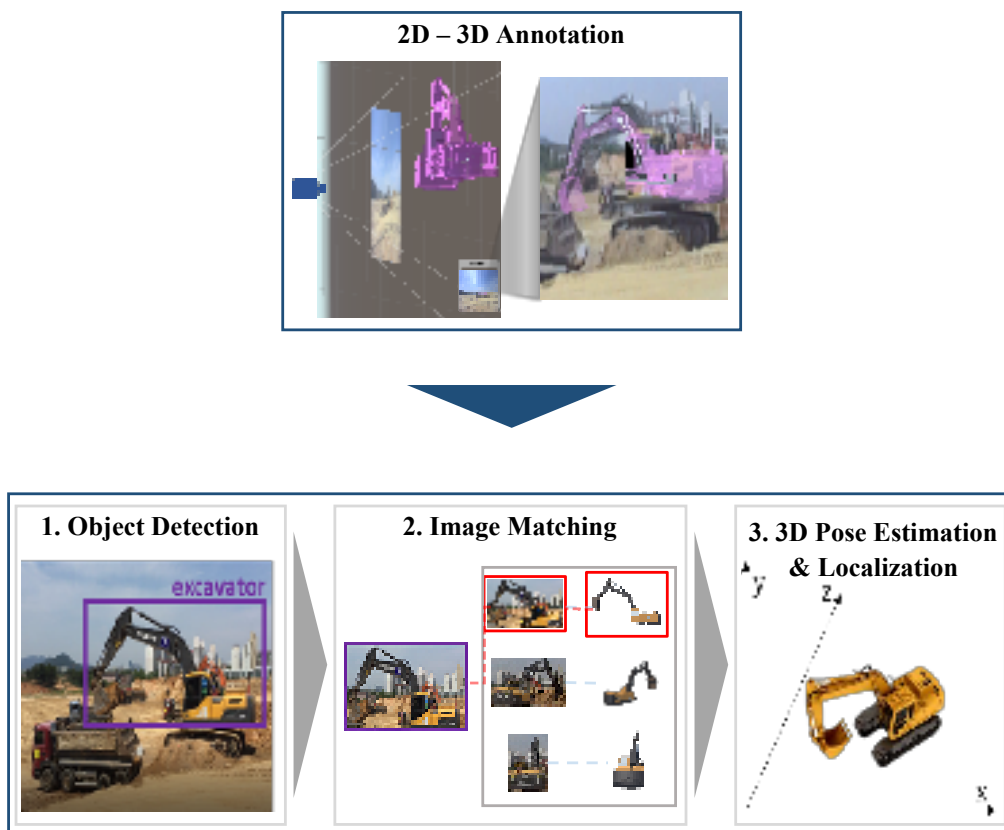


Figure 1.2 Overview of research methodology.

1.4 Dissertation Outline

This dissertation is composed of five chapters and the details for each chapters are as below.

Chapter 1. Introduction: This chapter covers the background and problems of the research, the goals and scopes of the research.

Chapter 2. Theoretical Background and Related Work: This chapter deals with how image analysis technology is applied for construction site monitoring, what it is intended to provide through it, and how studies are being conducted to estimate the pose of construction equipment.

Chapter 3. Construction Equipment's 3D Pose Estimation and Localization: This chapter describes the process performed for the 3D pose estimation and localization of construction equipment in the construction site image data. It consists of two parts: 2D - 3D annotation and development of pose estimation method.

Chapter 4. Experimental Results and Discussions: This chapter covers the result of applying the trained object detection model to test datasets, image matching results for extracted images, and the results of pose estimation and visualization. In addition, validation of the estimated pose information is performed by applying the root mean square error (RMSE), which is widely used as an evaluation index in pose estimation researches.

Chapter 5. Conclusions: This chapter summarizes achievements, contributions and limitations of this research, and describe the contents of future study.

Chapter 2. Theoretical Background and Related Work

This chapter describe the basic concepts of what is and why do construction site monitoring, and the various changes that the development of image analysis technology has brought to the research about construction site monitoring. It also investigates how the majority of existing studies on construction site monitoring carried out the identification, tracking, and classification of construction equipment. After that, describe how the previous studies to estimate the 3D pose of the equipment based on the image have been conducted.

2.1 Vision-based Construction Site Monitoring

Construction site monitoring is not only a necessary task to track ongoing construction work and provide up-to-date information, but also a critical component of construction project stakeholders (i.e., owners, contractors, etc.). In the past, a person in charge visited the site to determine the working status and time of the equipment, conduct a survey with the equipment operator, or monitor CCTV footage installed on the site in the control room. However, research on automatic site monitoring of vision-based is carried out along with the development of image analysis technology.

Dimitrov and Golparvar-Fard (2014) developed an vision-based material recognition algorithm using a support vector machine (SVM) that can classify materials in a single image for automatic construction progress monitoring and 3D modeling. Choi et al. (2008) uses stereo vision system and 3D CAD data to present a 3D object recognition framework and preliminary experimental results for automatic project progress monitoring. Wang et al. (2020) presents a vision-based framework incorporating computer vision methods such as mask region-based convolutional neural network (Mask R-CNN) and DeepSORT for automatic monitoring of precast wall construction progress. Deng et al. (2019) uses SVM to identify tiles in the image and combines camera location and tile boundary coordinate information with building information modeling (BIM) model to provide a method for automatic monitoring of interior tile work. Azar and McCabe (2012) present an object recognition framework using Haar-histogram of orientated gradient

(HOG) and Blob-HOG, which combine images and video processing methods to automatically recognize dump trucks in images and distinguish them from other earthwork machines. Kim et al. (2017) presents a construction equipment detection method to distinguish loader, excavator, dump truck, concrete mixer truck, and road roller using the region-based fully convolutional network (R-FCN) model.

Furthermore, vision-based construction site monitoring provides basic information for successful construction projects by continuously analyzing the progress of current construction projects and analyzing site safety. Azar et al. (2013) presents a server-customer interaction tracker framework that recognizes excavators and dump trucks, and tracks loading cycles using computer vision algorithms such as HOG and mean-shift. Bugler et al. (2016) presents a method for estimating the productivity of earthmoving operations by combining two vision-based technologies (i.e., photogrammetry and video analysis) to measure the volume of soil and track the process of excavation and transport equipment. Kim et al. (2018) proposes an automatic productivity evaluation method for earthwork processes in tunnels by integrating vision-based situation reasoning and construction process simulation using convolutional neural network (CNN) model. Kim et al. (2018) developed a vision-based activity identification framework that focuses on the interaction between excavators and dump trucks that can identify earthwork operations and measure work cycles. Luo et al. (2018) presents a method for distinguishing the various activities (e.g., placing concrete, leveling land, transporting goods, etc.) performed by different objects (i.e., equipment, workers) using the faster region-based convolutional neural network (Faster

R-CNN) and the ResNet-50 model.

Regarding the safety management, safety is a top priority in the construction process, and many researchers focus on the application of computer vision technology for site safety management (Jiang et al. 2020). In construction site safety monitoring, the application of computer vision technology mainly analyzes collision risk monitoring, prevention of falls from height risk, wearing personal protective equipment, etc. Seo et al. (2015) presents future research directions in the field of computer vision-based safety and health monitoring from a technical and practical use perspective. Kim et al. (2016) provides a field safety assessment system for extracting spatial information about objects (i.e., workers, equipment) with gaussian miniature model (GMM) and evaluating the safety level of each object using fuzzy set theory to monitor crashes with moving objects. Fang et al. (2018) uses a CNN model to detect workers in the process of filing complaints and to develop algorithms to identify workers without harness to solve the falls from height problem, the main cause of construction injury and death. Guo et al. (2018) proposes a deep learning-based computer vision technology method to identify workers without safety helmet in image sequences taken with unmaanned aerial vehicle (UAV).

2.2 Pose Estimation in Construction

2.2.1 Internet of Things-based Approach

Estimating the pose of construction equipment is divided into two categories: internet of things (IoT)-based method and vision-based method. In the case of IoT-based construction equipment pose estimation, the operating part of the equipment is obtained using sensors such as the inertial measurement unit (IMU), the ultra-wide band (UWB), the wireless local area network (WLAN), the global positioning system (GPS), and the global navigation satellite system (GNSS). Pose information of equipment obtained through IoT sensors is used to estimate the posture of current equipment, remote control of construction equipment, and productivity analysis of equipment. Zhang et al. (2012) estimated the position of the crane boom using UWB to prevent potential collisions and help carry out the work in crane operations. Sun et al. (2017) suggests a method for estimating the posture of dozer blades using real time kinematic (RTK) GPS and IMU sensors. Kang et al. (2018) uses accelerometers and GPS sensors to estimate the position of the excavator bucket, thereby presenting a precision measuring technique for excavators that can automatically measure the level error between the planned and work surfaces. Lee et al. (2019) developed the excavator remote control system based on the pose information for each driving part of the excavator acquired using GPS sensor and IMU sensor. Rashid and Louis (2019) uses the supervised learning method (i.e., decision tree, k-nearest neighbor, artificial neural network) to identify the possibility of equipment activity (i.e., engine

off, idle, loading, moving) recognition using the pose information acquired using IoT sensors. Pradhananga and Teaser (2012) categorizes the load time, haul time, unload time, and turn time of the excavator using 10 GPS sensors and analyzes the work productivity of the excavator.

Furthermore, studies have been conducted to predict the trajectory of equipment and future changes in posture using the pose information of acquired construction equipment to avoid potential collision risks. Luo et al. (2020) predicts the future posture of the equipment by integrating the construction equipment's pose information and recognized work information obtained through GPS, IMU sensors and using recurrent neural network (RNN) and gated recurrent unit (GRU) models. Fand et al. (2016) developed a real time pro-active safety assessment framework that uses IMU sensors to reconstruct the crane's site specific pose information in real time to identify movement and alert workers in advance of potential crash accidents. In addition, Vahdatikhaki and Hammad (2015) create a real time working space for equipment to prevent potential collision by taking into account the pose of equipment, state geometry and the speed characteristics of the equipment based on real time location system (RTLS) data acquired through IoT sensors.

However, in the case of IoT-based pose estimation, there are limitations such as IMU sensor error problems caused by sensor installation location or magnetic disturbance, and GPS sensor error caused by signal blockage and multipath, so studies are underway to correct this error. Tang et al. (2020) investigates the effect of pose estimation by the various installation position of IMU sensors to solve the error problem caused by the IMU sensor installation position. Pentek et al. (2017) developed an algorithm that corrects

the pose information of the estimated excavator arm regardless of the sensor installation location using IMU sensor with 3D gyroscopes and accelerometers. Vahdatikhaki et al. (2015) proposes an optimization-based method that uses the shape and operating characteristics of the excavator to set constraints on changes in the pose of the equipment and minimize the required amount of correction to improve the quality of pose estimation to improve the accuracy of the pose information generated by RTLS.

2.2.2 Vision-based Approach

Along with the development of image analysis technology, vision-based construction equipment pose estimation which estimates the pose of equipment by image alone without have to attach IoT sensors (e.g., GPS, IMU sensors, etc.) to each driving part of each equipment, has recieved substantial attention from many researchers. Azar et al. (2015) presents a computer vision-based framework using markers for real time pose estimating the boom and arm of excavator. Lundeen et al. (2016) attaches a 57cm (22.5 in.) marker to the joint of the excavator and simultaneously recognizes the marker and the landmark point in the image to estimate the position of the excavator. Yuan et al. (2017) uses stereo cameras to detect and track excavators, and uses fast directional chamfer matching algorithm to connect excavator key nodes and estimate poses through triangulation.

Furthermore, based on estimated pose information, real-time field monitoring, equipment work analysis, productivity and safety analysis were performed. Liang et al. (2018) uses scale-invariant feature transform (SIFT) and viewpoint feature histogram (VFH) method to estimate the location and direction of the equipment and to update the site layout in real time by synchronizing it in the BIM database. Roberts and Golparvar-Fard (2019) recognizes the behavior of the equipment and analyzes productivity by analyzing the interaction between the excavator and dump truck based on CNN. Souma-Gyimah et al. (2019) uses the single shot multi-box detector algorithm to estimate bucket's pose and combine CNN models to develop a

multipurpose vision model that prevents collisions in excavation work by recognizing large rocks.

In addition, research on estimating 3D pose of equipment was conducted using deep learning techniques without separate markers or IoT sensors. Liang et al. (2019) proposes a markerless 2D and 3D pose estimation method for articulated construction robots using state-of-the-art human pose estimation deep convolutional network (i.e., stacked hourglass network). Luo et al. (2020) developed a methodology framework to automatically estimate the 2D poses of construction equipment keypoints using three deep learning networks (i.e., stacked hourglass network, cascaded pyramid network, stacked hourglass network + cascaded pyramid network). Soltani et al. (2017) part-based detection of the excavator into three parts of the dipper, boom, and body, and extraction of skeleton to derive 2D bone of the excavator and 2D pose information of the key points.

Chapter 3. Construction Equipment's 3D Pose Estimation and Localization

This chapter describes the data annotation and method development for estimating and localizing the 3D pose of construction equipment in single camera's image data. In this research, the 3D pose estimation and localization of construction equipment consists of two parts: 2D - 3D annotation and development of pose estimation method.

2D - 3D annotation is the process of storing 3D pose information of construction equipment in the image data by performing the annotation between virtual models of the construction equipment (i.e., excavator, dump truck) and construction site image data (2D image).

Development of pose estimation method is divided into three steps: object detection, image matching, pose estimation. Step 1 object detection develops an object detection model that detects construction equipment (i.e., excavator, dump trucks) within the image by utilizing the construction site image data. And establish a crop image database for step 2, image matching. Step 2 image matching is a stage in which equipment is detected and extracted from construction site image data using trained object detection model (step 1) and image matching algorithm is used to derive the most similar image from the crop image database in step 1. Step 3 pose estimation is a stage in which 3D pose of construction equipment in input image data are estimated and

localized using 3D pose information of construction equipment stored in the image derived through image matching in step 2.

3.1 2D - 3D Annotation

2D - 3D annotation is the process of storing 3D pose information of construction equipment in the image data by performing the annotation between the 3D virtual model of the construction equipment and construction site image data.

In this research, a total of 90,000 construction site image data were collected, each of 30,000 for three construction sites, and annotation is performed between collected image and 3D virtual models of excavator and dump truck. The 3D virtual models for annotation are as shown below in Figure 3.1.

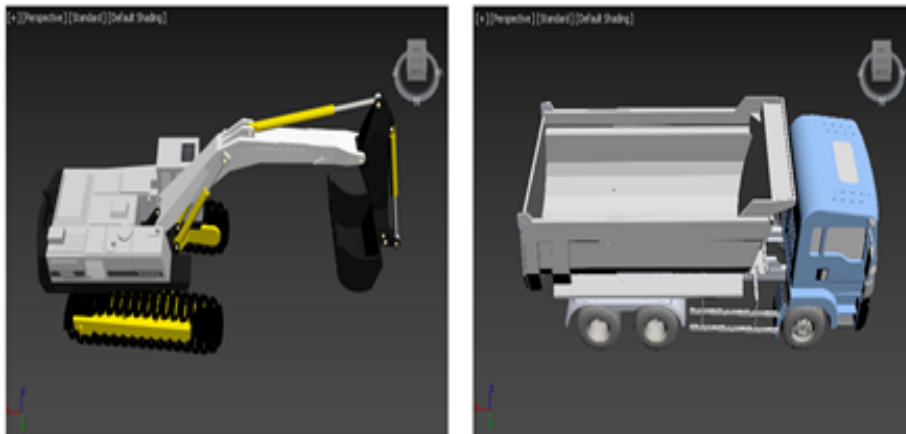


Figure 3.1 3D virtual models of excavator and dump truck.

Keypoints were specified in the 3D virtual models of the excavator and dump truck for estimation and visualization of the poses, as shown in Figures Figure 3.2 and Figure 3.3. And 3D pose information (x, y, z coordinate) for each keypoint was stored through annotation process.

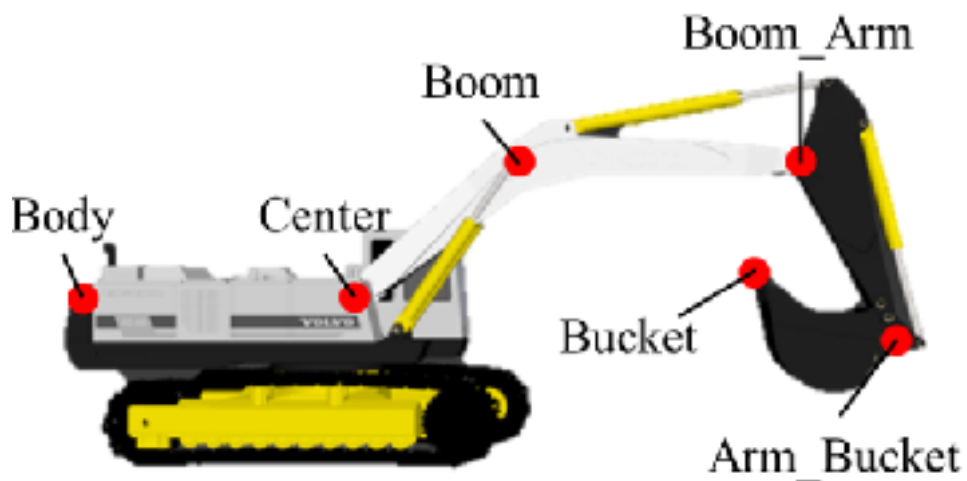


Figure 3.2 Keypoints of excavator 3D model.

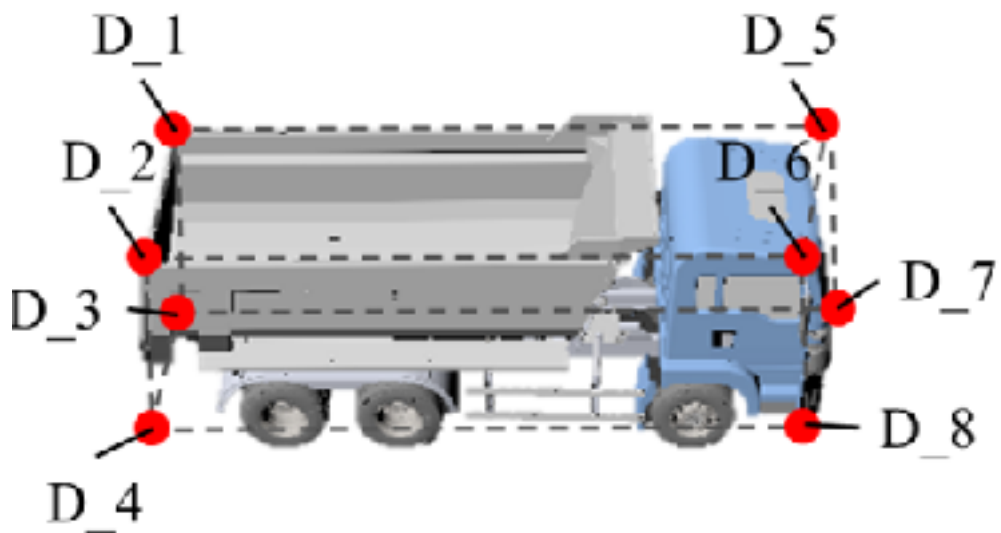


Figure 3.3 Keypoints of dump truck 3D model.

Unity game engine was used as a tool to perform annotation between 2D images and virtual models. Unity is a cross-platform game engine and is a tool used for 3D, virtual reality, and simulation in architecture, engineering and construction as well as game production. In this tool, the user can obtain a 2D image that is output through the camera on which the 3D object is placed by placing the camera and 3D object as shown in Figure 3.4. In addition, it is possible to manipulate a 3D object by writing and mounting a C# script so that the object shows the movement desired by the user, and obtain an image (i.e., 2D image of 3D object) when viewed with a camera.

In this research, a camera and an image plane are placed on the Unity game space to output the same as the construction site image as shown in Figure 3.4, and a 3D virtual model is placed on it. After that, a C# script was created to link with the kinematic parameters of each material, and the 3D model was overlaid on the 2D image by controlling the 3D model of the excavator and the dump truck.

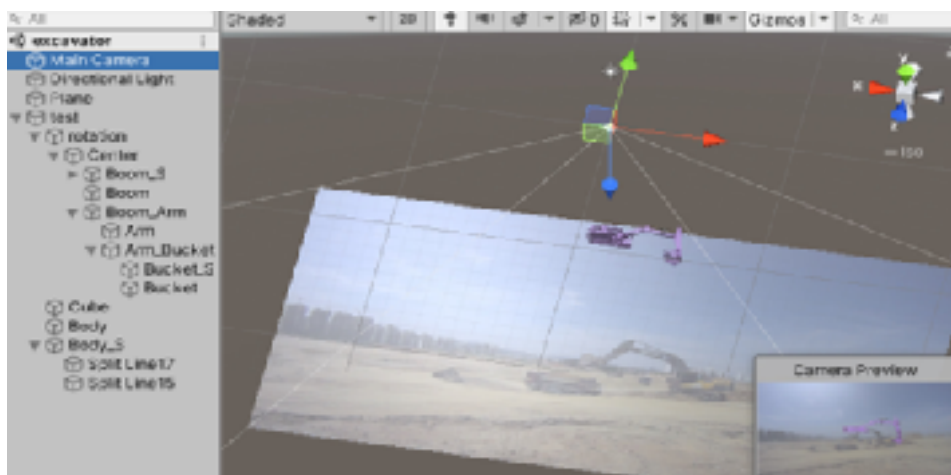
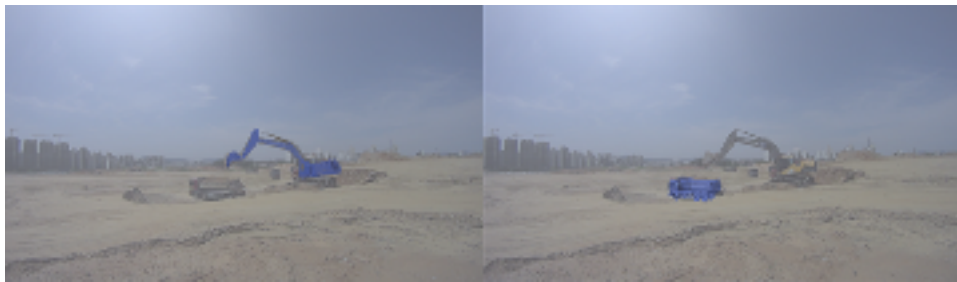


Figure 3.4 Unity game engine screen.

Using the Unity game engine, 3D virtual models of excavator and dump truck were annotated with construction image data collected at three construction sites. Examples are shown in Figure 3.5.



(a)



(b)



(c)

Figure 3.5 2D - 3D annotation examples. (a) Site 1, (b) Site 2, (c) Site 3.

In addition, an example of pose information for each keypoint of construction equipment stored through annotation is shown in Table 3.1 below.

Table 3.1 Stored pose information example by 2D - 3D annotation (cm).

Image No.	Body_x	Body_y	Body_z	Center_x
1	519.99	780.03	-133.67	390.55
2	519.81	779.80	-133.64	390.62
3	519.63	779.56	-133.60	390.69
4	519.45	779.32	-133.57	390.76
5	519.15	779.00	-133.52	390.88
6	518.85	778.67	-133.48	391.00
Image No.	Center_y	Center_z	Boom_x	Boom_y
1	879.27	-153.66	329.55	941.36
2	879.37	-153.67	329.75	941.60
3	879.47	-153.69	329.95	941.85
4	879.57	-153.70	330.14	942.09
5	879.71	-153.72	330.45	942.44
6	879.86	-153.74	330.76	942.79
Image No.	Boom_z	Boom_Arm_x	Boom_Arm_y	Boom_Arm_z
1	-52.77	191.93	1053.29	-27.83
2	-52.80	192.40	1053.89	-27.87
3	-52.82	192.86	1054.48	-27.91
4	-52.84	193.33	1055.08	-27.95
5	-52.89	194.08	1055.93	-28.07
6	-52.93	194.84	1056.78	-28.18
⋮				
⋮				
⋮				

3.2 Development of Pose Estimation Method

3.2.1 Object Detection

In this step, an object detection model is developed that uses construction site image data to detect construction equipment (i.e. excavator, dump truck) within the image. And establish a crop image database for next step, image matching.

Object detection is a widely used technique in computer vision, which classifies and localizes the types of objects that exist within image data. In this research, the faster region-based convolutional neural network (Faster R-CNN) model was used as an object detection model.

The Faster R-CNN model is a structure that adds a region proposal network (RPN) that generates region of interest (ROI) between the convolutional feature map and the ROI pooling layer in the Fast R-CNN model, which is widely used in the computer vision field. In other words, to solve the computational load problem that occurs in the ROI generation stage of the Fast R-CNN model, the concept of the region proposal network is added and the GPU is used. The structure is shown in Figure 3.6.

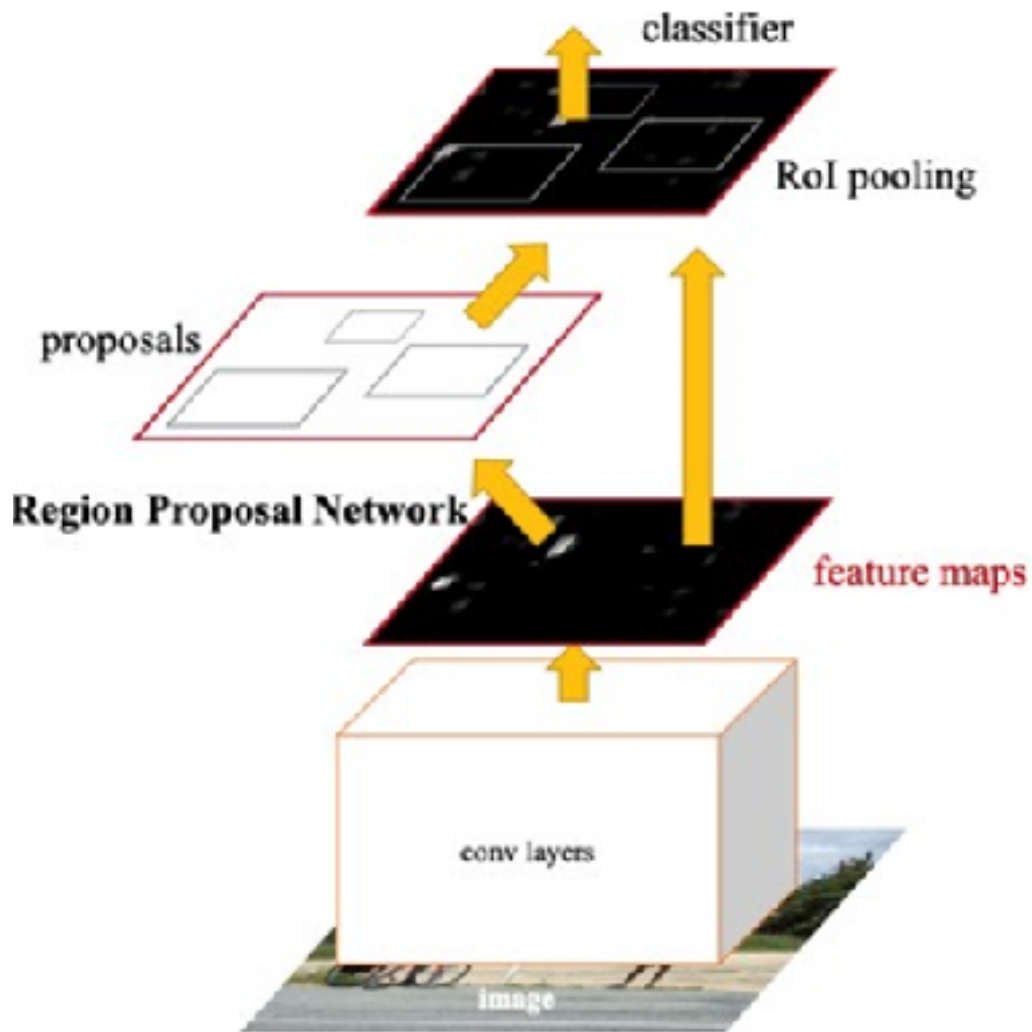


Figure 3.6 Faster region-based convolutional neural network model.

The object detection model was trained separately for each of the three sites, and two objects (i.e., excavator, dump truck) were trained. For model training, 30,000 images were collected for each site and 21,000 images were used as train dataset and 9,000 images (30%) as test dataset. Examples of train image and test image sets for each site are as shown in Figure 3.7.



(a)



(b)



(c)

Figure 3.7 Examples of train image and test image sets for each site.

(a) Site 1, (b) Site 2, (c) Site 3.

The crop image database for each construction equipment is constructed using the bounding box coordinates of each construction equipment (i.e., excavator, dump truck) that was labeled in the train images used for training the object detection model. When constructing the crop image database for each equipment, data was collected except for images in which no construction equipment was observed from 21,000 train images for each site. The number of extracted image data for each construction site and examples (Figure 3.8, 3.9, 3.10) are as follows.

1. Site 1: excavator - 21,000 images, dump truck - 21,000 images
2. Site 2: excavator - 21,000 images, dump truck - 20,791 images
3. Site 3: excavator - 21,000 images, dump truck - 9,128 images



Figure 3.8 Crop image database example (Site 1).



Figure 3.9 Crop image database example (Site 2).

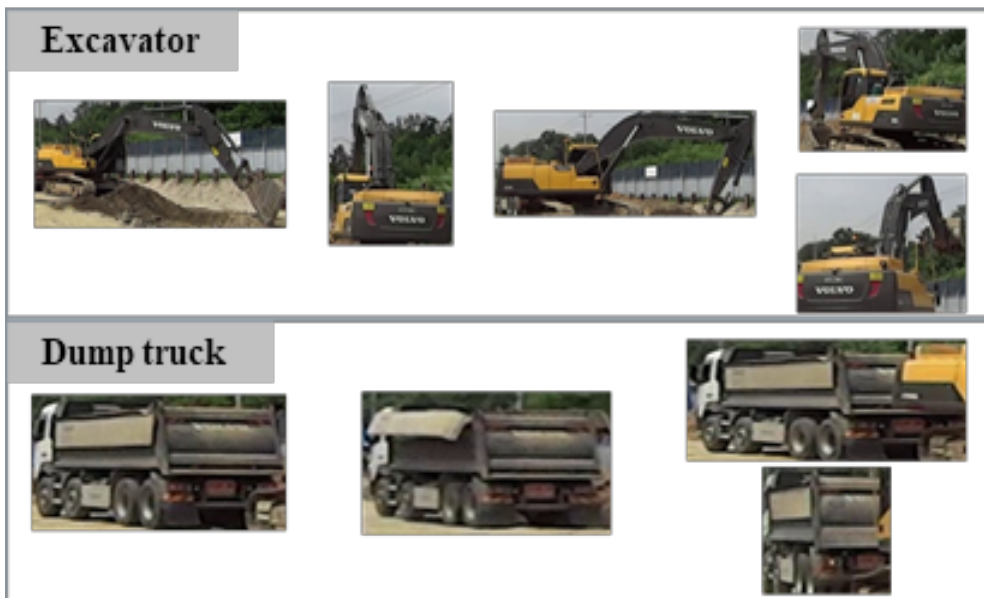


Figure 3.10 Crop image database example (Site 3).

3.2.2 Image Matching

Image matching is a stage in which equipment is detected and extracted from construction site image data using trained object detection model (step 1) and image matching algorithm is used to derive the most similar image from the crop image database (step 1).

Image matching used template matching algorithm provided by open source computer vision library (OpenCV). Template matching is an algorithm that finds the location of a specific image in an input image, and is a function that compares pixel values while sliding a specific image (i.e., template image) in a smaller area than the input image as shown in Figure 3.11.

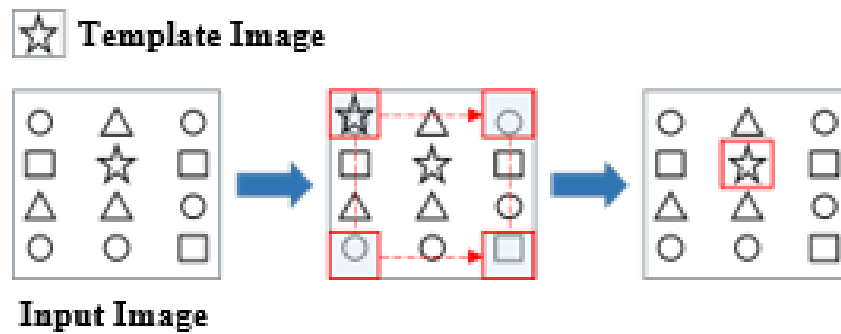


Figure 3.11 Template matching algorithm.

The equation used to compare pixel values between images in the template matching algorithm is as in Eq 1.

$$\mathbf{R}(x, y) = \frac{\sum_{x', y'} (\mathbf{T}(x', y') \cdot \mathbf{I}(x + x', y + y'))}{\sqrt{\sum_{x', y'} \mathbf{T}(x', y')^2 \cdot \sum_{x', y'} \mathbf{I}(x + x', y + y')^2}} \quad \mathbf{Eq.1}$$

Input image (**I**) : Image that is expected to match the template image

Template image (**T**) : Image which will be compared to the input image

Result matrix (**R**) : Template matching result value in input image pixel position at (x, y)

(x, y) means pixel location in input image and (x', y') means pixel location in template image. To identify the matching area, the template image is slid and compared to the input image such as Figure 3.11. Sliding means moving template image one pixel at a time (left to right, top to bottom). While sliding the template image on the input image, the pixel values of the image are calculated as a matrix for each pixel position, and this is stored in the result matrix. After sliding for all pixels in the input image to obtain values, the location of the pixel with the highest result value is found, and this is derived as the result value of the template matching.

Template matching algorithm is an algorithm that finds the position of a specific image in the input image as described above. However, the purpose of performing image matching in this research is not to find where the detected construction equipment is located within the input image data, but to

find the most similar to detected construction equipment image in the crop image database. That is, comparing different images to derive similarity. To this end, in this research, the detected construction equipment image was set as input image and the crop image database was set as template image in the template matching algorithm. An algorithm for image matching was created by adding two processes: selecting comparative image group and resizing the comparative image.

In the comparative image group selection phase, only images in the crop image database that have an aspect ratio similar (e.g., within ± 0.2) to the aspect ratio of the detected construction equipment image are selected as the image matching comparison target group in order to improve the performance of image matching.

In the comparison image resize phase, image matching was performed by resizing the size of the comparison target group images to match the size of the detected construction equipment image.

The developed image matching algorithm is shown in Figure 3.12.

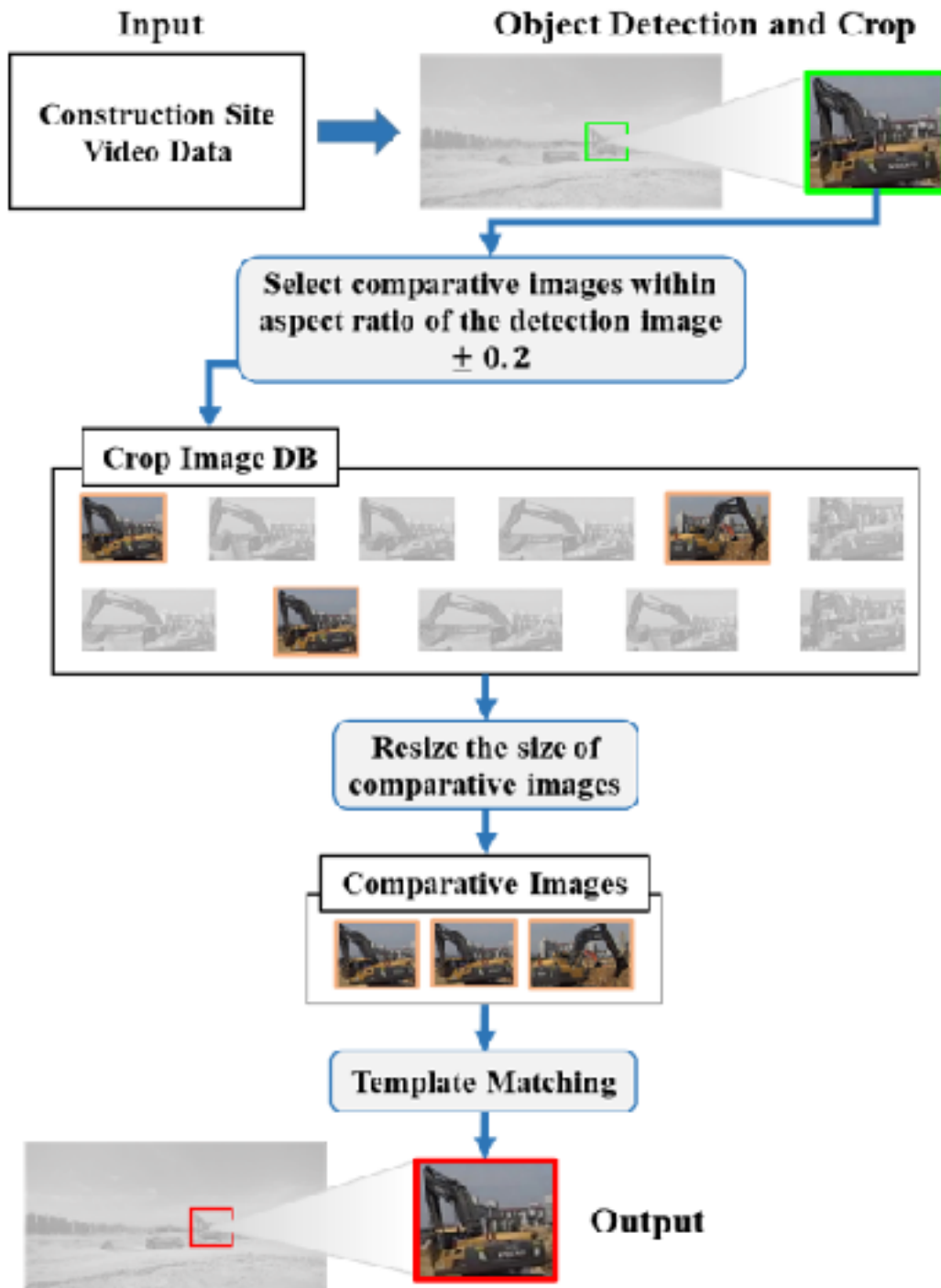


Figure 3.12 Image matching algorithm.

3.2.3 Pose Estimation and Localization

Pose estimation is a stage in which 3D pose of construction equipment in input image data are estimated and localized using 3D pose information of construction equipment stored in the image derived through image matching.

In this research, construction site image data were collected to train the object detection model, and a crop image database was built for image matching. After that, 3D virtual models and annotations were performed on construction equipment (i.e., excavator, dump truck) in the image data collected in Chapter 3.1 2D - 3D Annotation. Therefore, 3D pose (x, y, z) information for each keypoint of construction equipment is stored in the image derived from the crop image database through Chapter 3.2.2 image matching, as shown in Figure 3.13.

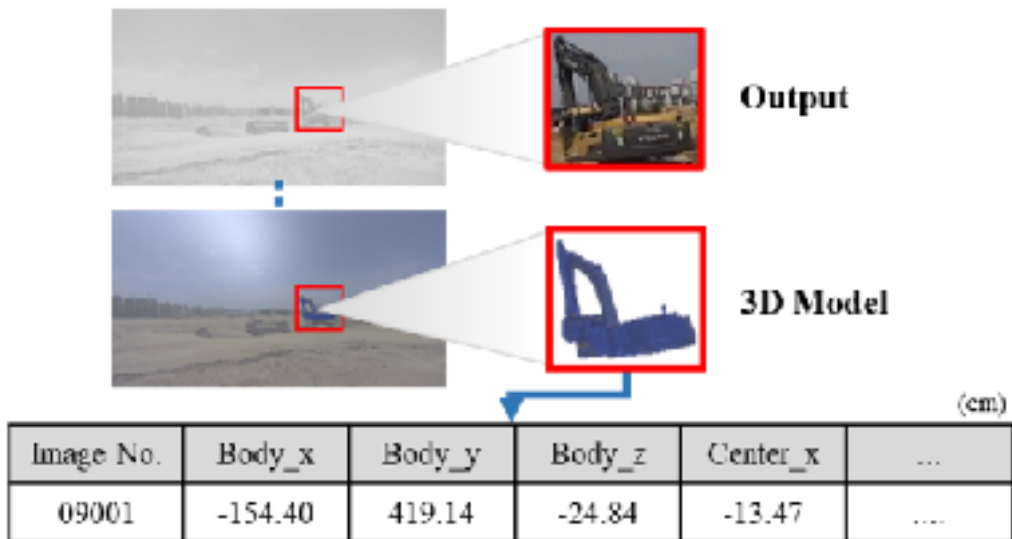


Figure 3.13 Example of pose information stored in the image matching result.

However, the image matching algorithm suggested in chapter 3.2.2 performs comparison between extracted images of the part of construction equipment detected through the object detection model in the construction site image data. In other words, when image matching is performed, the location of the detected construction equipment (i.e., coordinates of the detected construction equipment's bounding box) cannot be considered. If the images of the detected construction equipment are identical such as Figure 3.14, then the same pose will be taken regardless of where the construction equipment is located in the whole image.

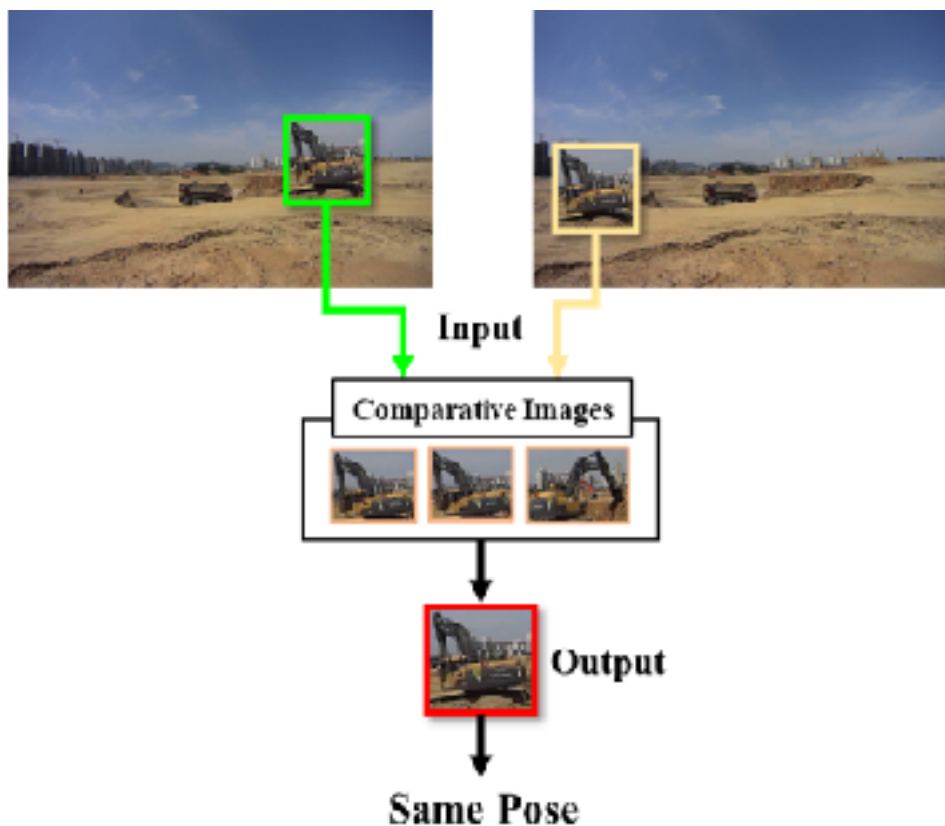


Figure 3.14 Same pose will be taken regardless of crop image's location.

In addition, before applying the template matching algorithm in the image matching algorithm, the process of matching the size of the comparison target image and the detected construction equipment image (i.e., input image) was performed. For this reason, if the comparison target image and the detected construction equipment image have the same shape, the pose information stored in the comparison target image (i.e., in crop image DB) is taken as it is, even if the size is different. That is, the pose information stored in the image derived through the image matching algorithm does not reflect the location information and the size information that the input image has in the detected original image. Therefore, additional post-processing processes are required to estimate the 3D pose of the equipment in the input image data using the 3D pose information of the construction equipment stored in the image derived through image matching.

In this research, post-processing was performed by using the camera arrangement in the Unity game engine used for annotation process, the distance to the plane, the coordinates of the construction equipment 3D model and the bounding box coordinates of detected object, and the similarity theorem of the triangle. Similarity in a figure means that when two figures are given, one side is reduced or enlarged at a reduction ratio to become congruent with the other.

The conditions for satisfying this similarity in a triangle are as follows:

1) If two triangles have two of their angles equal, the triangles are similar (AA similarity).

2) If two triangles have two pairs of sides in the same ratio and the included angles are also equal, then the triangles are similar (SAS similarity).

3) If two triangles have three pairs of sides in the same ratio, then the triangles are similar (SSS similarity).

In this research, post-processing was performed using the concept of “If two triangles have two of their angles equal, the triangles are similar (AA Similarity)”, which is one of the similarity conditions of triangles as shown in Figure 3.15.

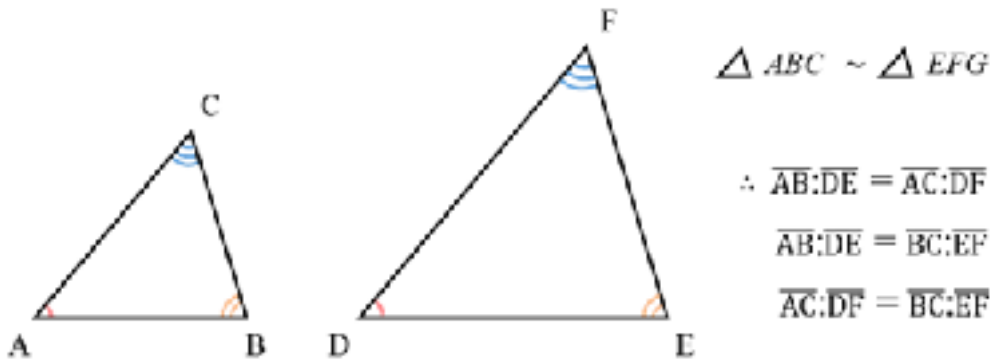


Figure 3.15 Triangle AA similarity.

The details of post-processing performed to estimate the 3D pose of construction equipment in the input image using this condition are as follows.

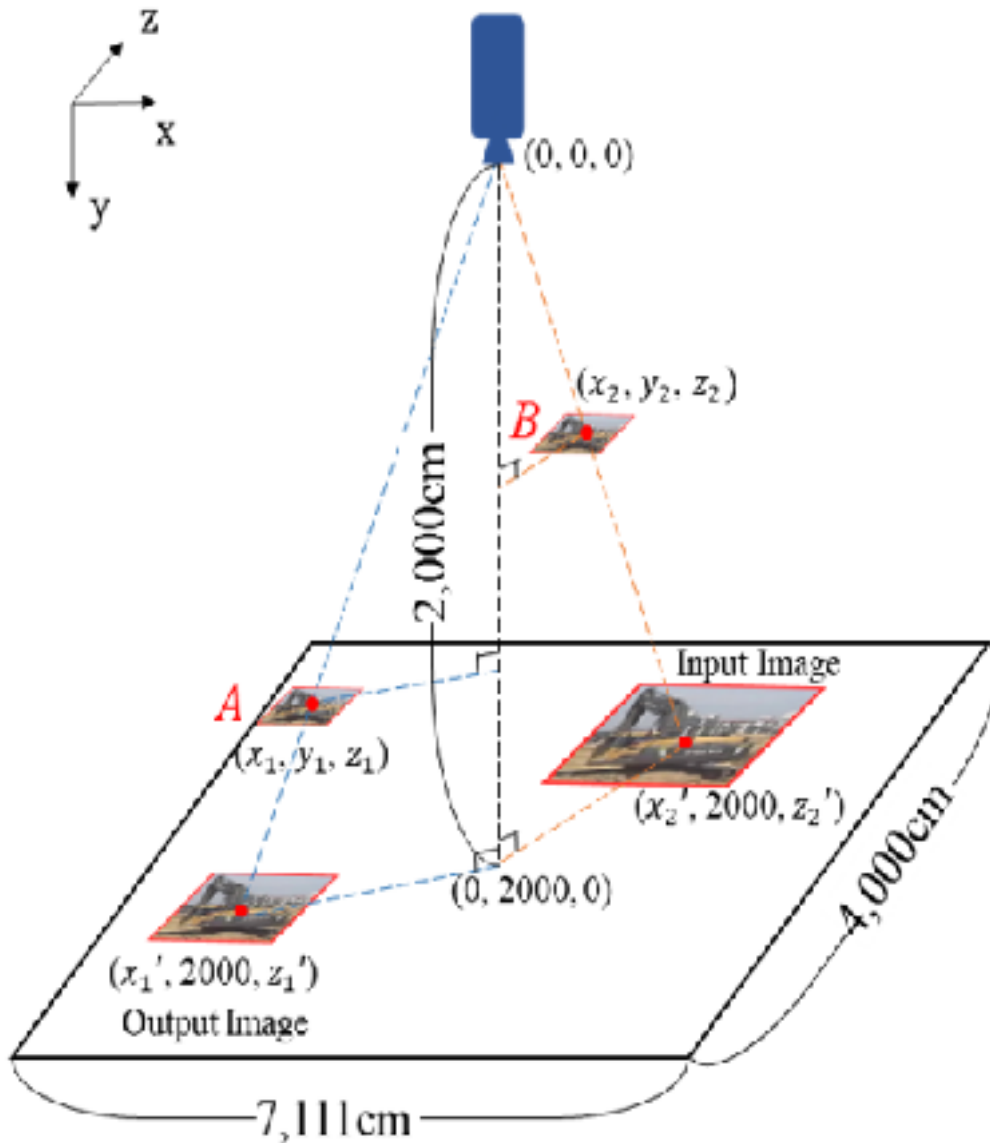


Figure 3.16 Layout in Unity game engine space.

As shown in Figure 3.16, when annotation between the 2D image and the 3D virtual model was performed, the coordinate of the camera in the Unity space was set to (0, 0, 0) and the distance to the camera background plane was

fixed to 2,000cm. In addition, in order to keep the image output from Unity camera at the same scale (i.e., width: height = 16:9) as the original 2D image, the width of the plane was set to 7,111cm and the height of the plane was 4,000cm considering the distance between the plane and the camera and the angle of view of the camera. In Figure 3.16, the y_1 value of (x_1, y_1, z_1) is the y-coordinate value of the Boom_Body, which is the rotation axis of the excavator body, among the keypoints of the excavator 3D virtual model saved in the annotation process (in the case of dump truck, it is the y coordinate value of the center point of the object). A is a projection of an annotated 3D virtual model across the y_1 coordinates and parallel to the plane. Also, x_1 and z_1 denote the midpoint coordinates of A. B is a projection of the y-coordinate value of the Boom_Body (in case of dump truck, it is the y coordinate value of the center point of the object) in the input image and parallel to the plane so that it has the same size as A. That is, A denotes 3D information of the construction equipment 3D model stored in the output image of image matching. B denotes 3D information of the construction equipment in the input image in image matching. Therefore, by adding the center point's coordinate difference of B and A $(x_2-x_1, y_2-y_1, z_2-z_1)$ from the 3D pose value of each keypoint of the construction equipment stored in the image derived through image matching, 3D information estimation is performed that reflects the location information of the construction equipment in input image and its size.

The value of y_1 is data stored through the annotation process in the y-coordinate of the Boom_Body (the y coordinate of the center in the case of a dump truck), and the remaining 5 unknowns (x_1, x_2, y_2, z_1, z_2). A detailed description of the process is as follow.

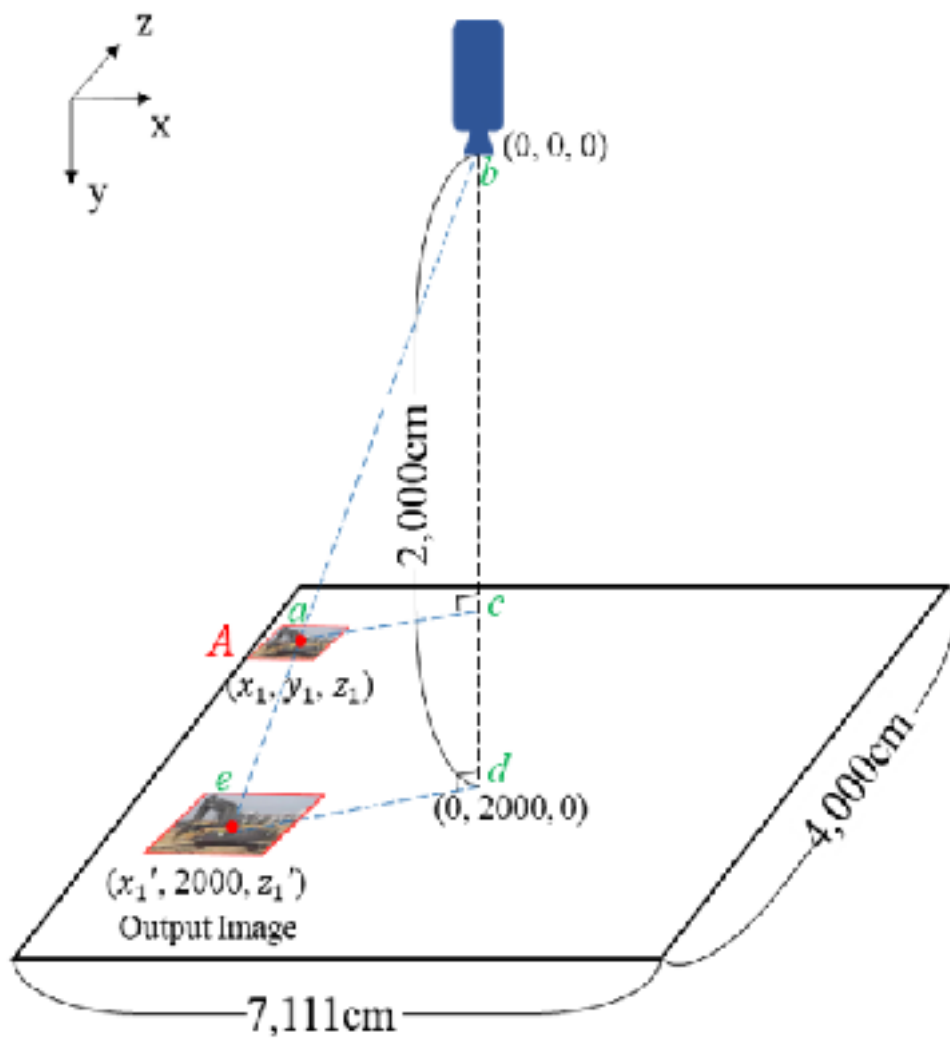


Figure 3.17 AA similarity condition to calculate (x_1, y_1, z_1) .

First of all, as shown in Figure 3.17, Δabc and Δebd are $\angle abc = \angle ebd$ and $\angle acb = \angle edb$, so it satisfies one of the triangular similarity conditions, “If two triangles have two of their angles equal, the triangles are similar (AA Similarity).”. Therefore, $\overline{bc}:\overline{bd}=\overline{ac}:\overline{ed}$. It is also already known that the length of \overline{bd} is 2,000cm, and the length of \overline{ed} can be obtained at $\sqrt{x_1'^2+z_1'^2}$.

x_1' and z_1' are replaced by the coordinate value of the construction equipment's bounding box, which has a value of 0 left bottom and 1 right top, and multiplied by the size of the plane (7,111cm x 4,000cm) set by the Unity game engine.

For example, when the coordinates of the center of the labeled bounding box are (-0.3, -0.1), the values of x_1' and z_1' are respectively

$$x_1' = -0.3 \times 7,111\text{cm} \cong -2,133\text{cm}, \text{ and}$$

$$z_1' = -0.1 \times 4,000\text{cm} = -400\text{cm}.$$

Also, (x_1, y_1, z_1) and $(x_1', 2000, z_1')$ are two points located on a single straight line with the same slope. Therefore, the values of x_1 and z_1 can be obtained by multiplying the values of x_1' and z_1' by the ratio of \overline{bd} and \overline{bc} , respectively.

For example, if $x_1' = -2,133\text{cm}$, $z_1' = -400\text{cm}$ and let $y_1 = 1,300\text{cm}$,

$$x_1 = -2,133\text{cm} \times \frac{\overline{bc}}{\overline{bd}} = -2,133\text{cm} \times \frac{1,300\text{cm}}{2,000\text{cm}} \cong -1,386\text{cm},$$

and

$$z_1 = -400\text{cm} \times \frac{\overline{bc}}{\overline{bd}} = -400\text{cm} \times \frac{1,300\text{cm}}{2,000\text{cm}} = -260\text{cm}.$$

In this way, the value of (x_1, y_1, z_1) can be obtained using the similarity of the triangle, and the value of (x_2, y_2, z_2) can be obtained in the same way by using the similarity of the triangle in Figure 3.18.

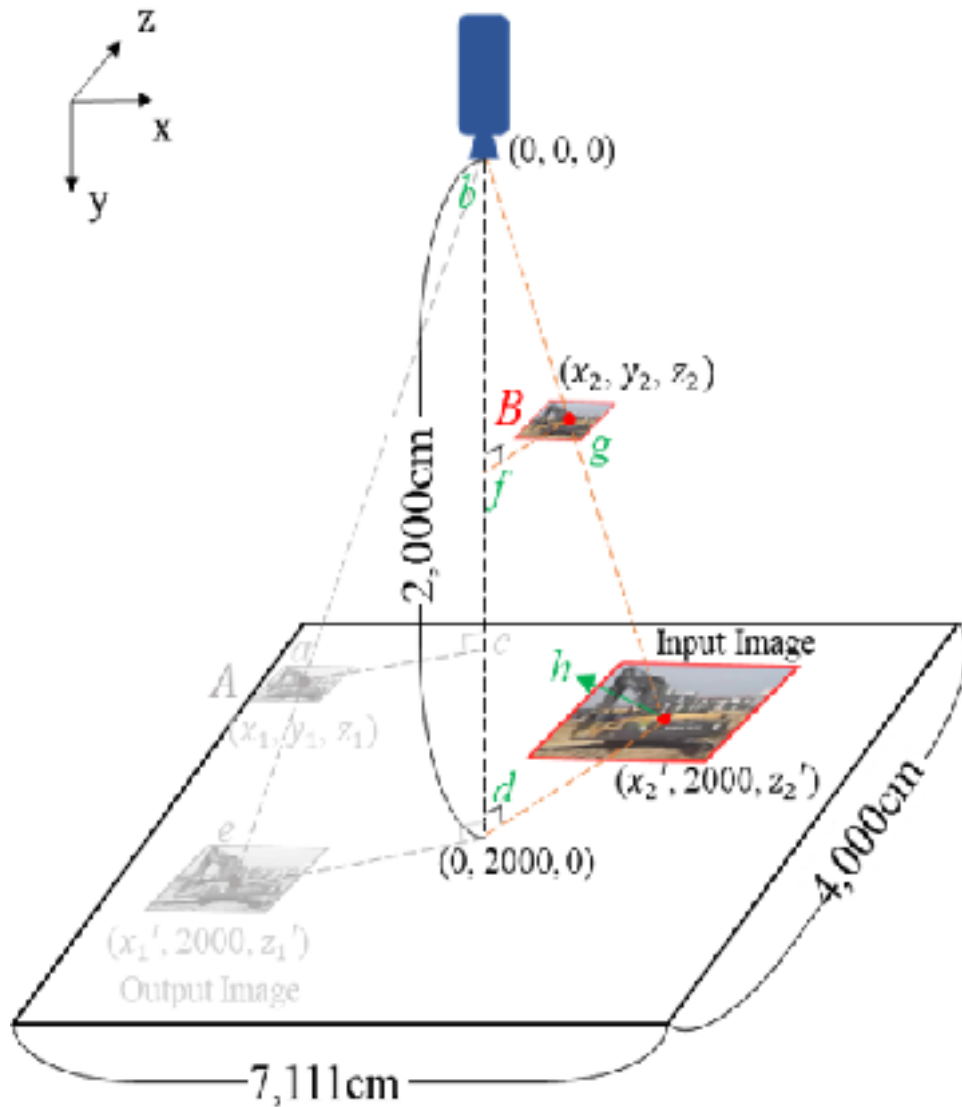


Figure 3.18 AA similarity condition to calculate (x_2, y_2, z_2) .

As show in Figure 3.18, since Δbfg and Δbdh are $\angle fbg = \angle dbh$ and $\angle bfg = \angle bdh$, it can be seen that AA similarity is satisfied among the similarity conditions of a triangle. And A and B have the same width.

Therefore, the value of (x_2, y_2, z_2) can be obtained by using the similarity of the triangle and the width of B as the process used to obtain the value of (x_1, y_1, z_1) . For example, width of B : width of input image = $\overline{bf} : \overline{bd}$. The rest of the process is the same.

Lastly, estimate the 3D pose of construction equipment in input image by adding the previously obtained the center point's coordinate difference of B and A $(x_2 - x_1, y_2 - y_1, z_2 - z_1)$ from the 3D pose value of each keypoint of construction equipment stored in the image derived through image matching.

After that, outliers among the 3D pose values of construction equipment in the input image estimated through image matching and post-processing are removed. The flowchart of the entire process of correcting by removing outliers from the estimated 3D pose value is shown in Figure 3.19.

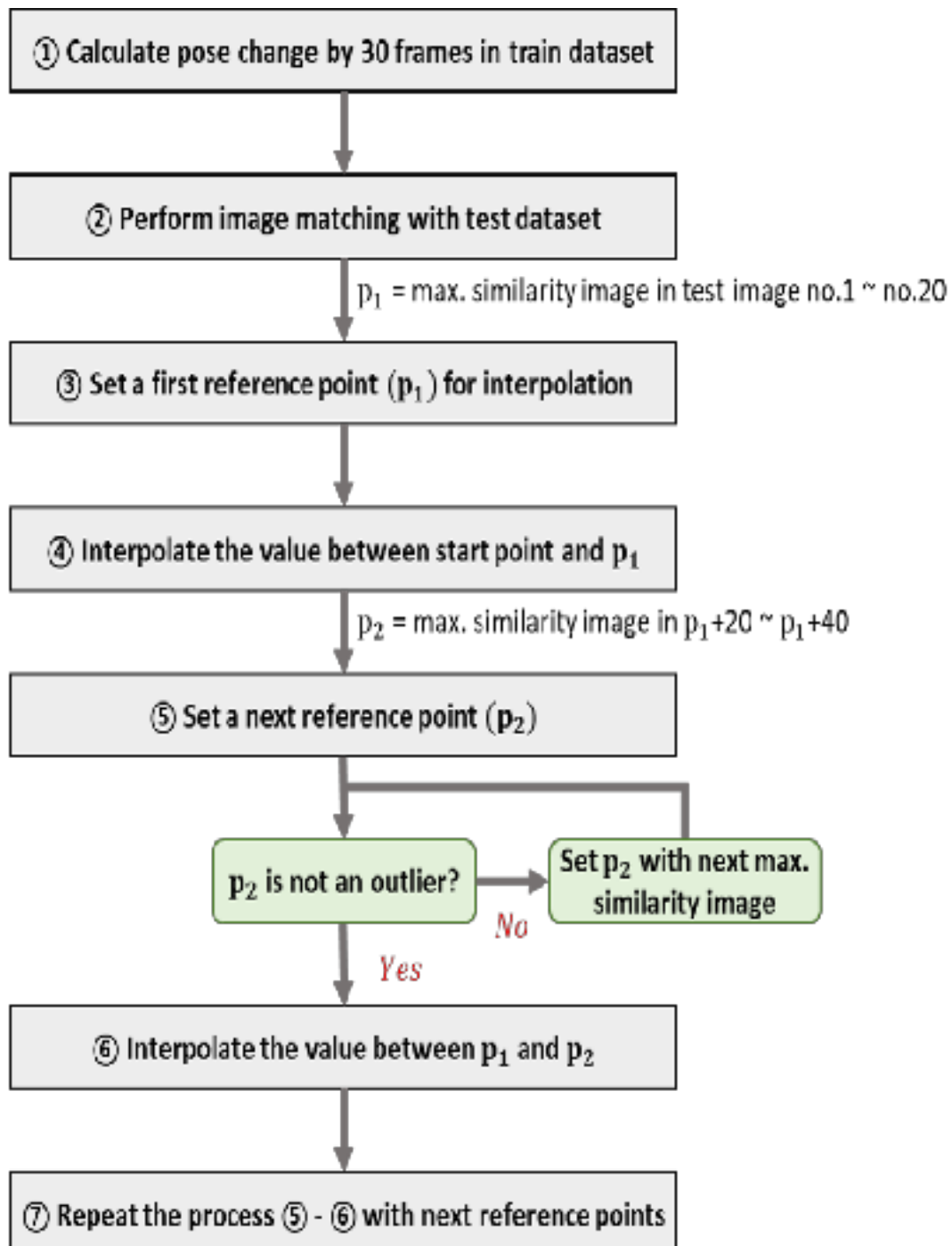


Figure 3.19 Flowchart of the entire process of correcting by removing outliers.

Before detailed description of each stage of the flowchart, assume two things:

1) The input image (i.e., test dataset) for pose estimation consists of a series of images.

2) Construction equipment moves similar speed in the same direction within 0.5 sec (i.e., 30 frames) without significant fluctuations.

To elaborate on each step, the "① Calculate pose change by 30 frames in train dataset" calculates the change during 0.5 sec (i.e., 30 frames) for the construction equipment pose values that were annotated to 21,000 train images for each site. Calculate the change in poses by 30 frames of the joint of the bucket and arm with the most variation in the pose (in the case of the dump truck, calculate the change in poses by 30 frames of the center of the object). The largest of the calculated pose change values for each 30 frames of the train dataset is set as the limit value that the pose change within 30 frames of construction equipment in the test dataset estimated using the pose estimation method.

In the step of "② Perform image matching with test dataset", a test dataset is composed of continuous images collected from image data of a construction site, and image matching is performed.

In the step of "③ Set a first reference point (p_1) for interpolation", set the first reference point to correct the estimated pose information. The first reference point is the image with the highest similarity value as a result of

template matching among test datasets entered as input images in the pose estimation model, from the first image to the 20th image (i.e., from test no.1 to test no.20).

In the step of “④ Interpolate the value between start point and p_1 ”, the value between test no.1 and the first reference point p_1 is obtained in step ③ is interpolated according to the second assumption “Construction equipment moves at a similar speed in the same direction within 0.5 sec (i.e., 30 frames) without significant fluctuations”. That is, the pose value of test no.n between test no. p_1 and test no.1 is calculated according to the equation below.

$$(\text{pose of}) \text{ no.n} = \text{no.1} + (\text{no.p}_1 - \text{no.1}) \times \frac{n}{p_1-1} \quad \text{Eq.2}$$

In the step of “⑤ Set a next reference point (p_2)”, find p_1+30 that moves in a similar direction and at a similar speed from p_1 according to the second assumption. And add ± 10 frames to based on p_1+30 . That is, among p_1+20 to p_1+40 , the image with the highest similarity value as a result of template matching is set to p_2 . Since the test dataset is a continuous image (assumption 1), it is judged as an outlier when the pose information changes too large when comparing the pose information of the previous and subsequent frames. Use the limit of the change in pose for 30 frames of construction equipment obtained from “① Calculate pose change by 30 frames in train dataset” as the reference value for determining the ideal value. The equation for this is as follows. Let maximum pose change for 30 frames in train dataset as L.

if (pose of) $no.p_2 - no.p_1 > L \times \frac{p_2 - p_1}{30}$, then p_2 is an outlier **Eq.3**

If the value of p_2 is higher than the limit value (that is, if it is determined as an outlier), the image with the second-order similarity value among the template matching results from p_1+20 to p_1+40 is set as p_2 . And outlier determination is performed again through Eq.3. This process proceeds until p_2 that satisfies Eq.3 is found from p_1+20 to p_1+40 . If all images between p_1+20 and p_1+40 are determined to be outlier, the image with the highest similarity among the template matching results from p_1+20 to p_1+40 is set to p_2 .

In the step of “⑥ Interpolate the value between p_1 and p_2 ”, the values of reference points p_1 and p_2 obtained earlier are interpolated with equal difference. And the same way as in step ④, interpolate the pose value of the equipment in the image between using Eq.2.

In the step of “⑦ Repeat the process ⑤ - ⑥ with next reference points”, the previous ⑤ - ⑥ process up to the last image of the test dataset, that is, the process of obtaining a reference point for interpolation and interpolating between them is repeatedly performed.

Chapter 4. Experimental Results and Discussions

This chapter covers the result of applying the trained object detection model (construction equipment detection and extraction) to 9,000 test datasets for each site collected at three construction sites, image matching results for extracted images, and the results of pose estimation and visualization. In addition, for validation of the construction equipment pose information estimated from test data set, the root means square error (RMSE) value which is widely used as an evaluation index in the pose estimation field is calculated.

4.1 Object Detection and Image Matching

In this research, image data was collected from three construction sites, and 30,000 images were extracted for each site. And 21,000 images (70%) were used as train dataset and 9,000 images (30%) were used as test dataset to train object detection models that detect 2 objects (i.e., excavator, dump truck). The performance (i.e., average precision, AP) and examples of trained object detection models are shown below (Table 4.1 and Figure 4.1, 4.2, 4.3).

Table 4.1 Performance of trained object detection model (AP).

	Site 1	Site 2	Site 3
Excavator	97.60%	98.32%	98.49%
Dump truck	92.01%	97.58%	95.38%

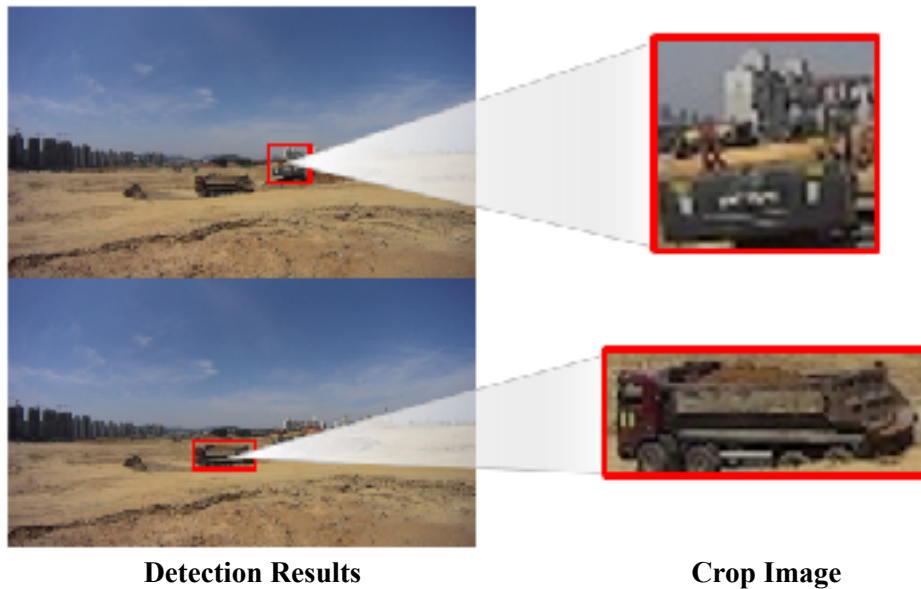


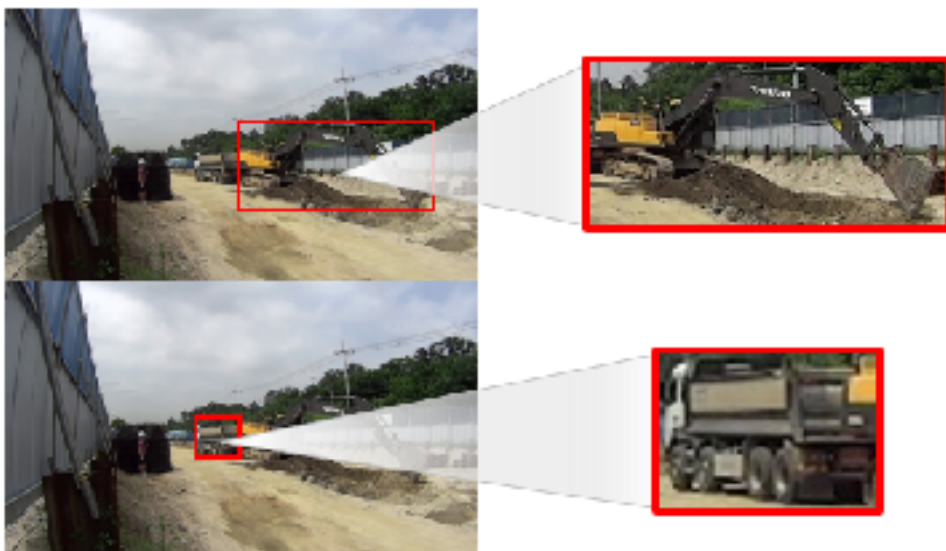
Figure 4.1 Examples of experimental results (Site 1).



Detection Results

Crop Image

Figure 4.2 Examples of experimental results (Site 2).



Detection Results

Crop Image

Figure 4.3 Examples of experimental results (Site 3).

In addition, using the bounding box coordinates of train dataset used for training object detection models, the crop image database for each construction equipment was established for image matching in 21,000 train images. The object detection model was used to detect and extract construction equipment from 9,000 test images and use it as a test dataset for image matching. Examples of performing image matching by site and construction equipment are shown in Figure 4.4 and 4.5.

Image matching similarity = 0.9725



Test no.7227



Crop image DB no.16667

(a)

Image matching similarity = 0.9917



Test no.8947



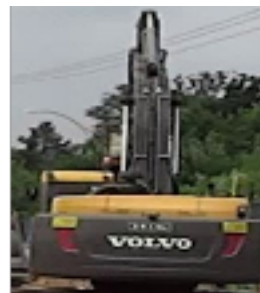
Crop image DB no.5583

(b)

Image matching similarity = 0.9883



Test no.6667



Crop image DB no.3507

(c)

Figure 4.4 Examples of image matching results (excavator).
(a) Site 1, (b) Site 2, (c) Site 3.

Image matching similarity = 0.9849



Test no.8544



Crop image DB no.3736

(a)

Image matching similarity = 0.9927



Test no.8704



Crop image DB no.701

(b)

Image matching similarity = 0.9813



Test no.5639



Crop image DB no.1026

(c)

Figure 4.5 Examples of image matching results (dump truck).
(a) Site 1, (b) Site 2, (c) Site 3.

As shown in Figures 4.4 and 4.5, it was confirmed that the result of image matching on the test dataset for each site and construction equipment showed quite good performance.

4.2 Pose Estimation and Localization

The 3D pose of the construction equipment in the test dataset is estimated and localized using the 3D pose information of the construction equipment stored in the image derived through image matching according to chapter 3.2.2 and chapter 3.2.3. The overall process of performing pose estimation and localization after image matching is shown in Figure 4.6.

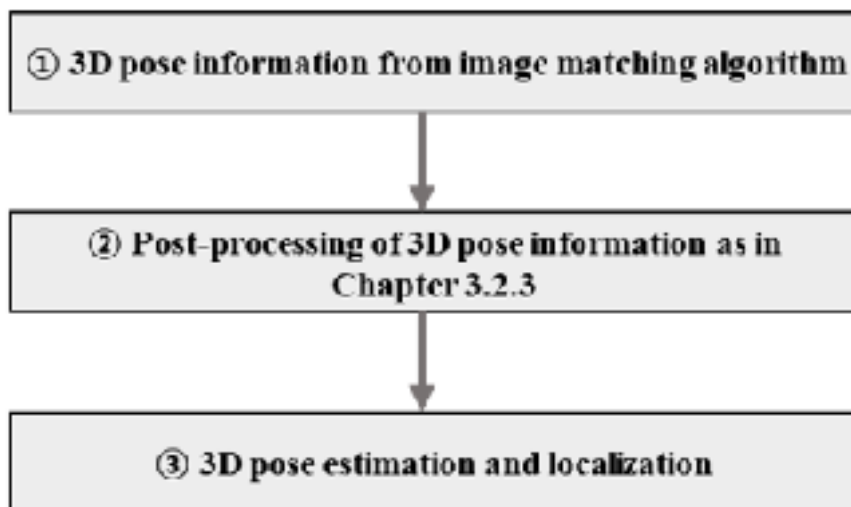


Figure 4.6 Process of performing pose estimation and localization.

First, before proceeding with the post-processing process (②) for pose estimation and localization, an example of pose information (①) of construction equipment in the test dataset derived by image matching is shown in Table 4.2.

Table 4.2 Example of pose estimation in the test dataset (Site 1, excavator).

Image No.	Similarity	Arm_Bucket_x	Arm_Bucket_y	Arm_Bucket_z
1	0.9288	315.63	634.02	91.78
2	0.9317	322.12	661.81	76.11
3	0.9285	315.63	634.02	91.78
4	0.9298	296.18	550.66	138.79
5	0.9302	296.18	550.66	138.79
6	0.9304	296.18	550.66	138.79
7	0.9346	302.67	578.45	123.12
8	0.9365	360.79	1162.71	-144.95
9	0.9353	296.18	550.66	138.79
10	0.9365	357.46	1162.39	-144.47
11	0.9315	309.15	606.23	107.45
12	0.9326	283.22	495.09	170.12
13	0.9350	263.77	411.73	217.13
14	0.9390	344.16	1161.09	-142.57
15	0.9357	344.16	1161.09	-142.57
16	0.9372	344.16	1161.09	-142.57
17	0.9351	344.16	1161.09	-142.57
18	0.9364	344.16	1161.09	-142.57
19	0.9395	344.16	1161.09	-142.57
20	0.9396	344.16	1161.09	-142.57

As a result of image matching, the image with the highest similarity (test image no.20 in this example) is set to p_1 , and interpolate values between no.1 and p_1 . The results are shown in Table 4.3.

Table 4.3 Pose interpolation results from no.1 to p_1 .

Image No.	Similarity	Arm_Bucket_x	Arm_Bucket_y	Arm_Bucket_z
1	0.9288	315.63	634.02	91.78
2	0.9317	317.13	661.76	79.45
3	0.9285	318.63	689.50	67.11
4	0.9298	320.13	717.24	54.78
5	0.9302	321.64	744.98	42.44
6	0.9304	323.14	772.72	30.11
7	0.9346	324.64	800.46	17.77
8	0.9365	326.14	828.20	5.44
9	0.9353	327.64	855.94	-6.89
10	0.9365	329.14	883.68	-19.23
11	0.9315	330.65	911.43	-31.56
12	0.9326	332.15	939.17	-43.90
13	0.935	333.65	966.91	-56.23
14	0.939	335.15	994.65	-68.56
15	0.9357	336.65	1022.39	-80.90
16	0.9372	338.15	1050.13	-93.23
17	0.9351	339.66	1077.87	-105.57
18	0.9364	341.16	1105.61	-117.90
19	0.9395	342.66	1133.35	-130.24
20	0.9396	344.16	1161.09	-142.57

Then, set the next reference point (p_2) among p_1+20 to p_1+40 (test image no.40 to no.60 in this example). Table 4.4 shows the pose estimation results between no.40 and no.60.

Table 4.4 Pose estimation results from no.40 to no.60 (Site 1, excavator).

Image No.	Similarity	Arm_Bucket_x	Arm_Bucket_y	Arm_Bucket_z
40	0.9441	281.86	1149.82	-132.20
41	0.9429	272.32	1146.30	-130.10
42	0.9439	272.32	1146.30	-130.10
43	0.9446	272.32	1146.30	-130.10
44	0.9436	272.32	1146.30	-130.10
45	0.9413	272.32	1146.30	-130.10
46	0.9427	275.50	1147.47	-130.80
47	0.9417	269.14	1145.12	-129.41
48	0.9417	272.32	1146.30	-130.10
49	0.9399	269.14	1145.12	-129.41
50	0.9393	269.14	1145.12	-129.41
51	0.9396	269.14	1145.12	-129.41
52	0.9385	269.14	1145.12	-129.41
53	0.9386	269.14	1145.12	-129.41
54	0.9396	288.22	1152.17	-133.60
55	0.9397	269.14	1145.12	-129.41
56	0.9384	278.68	1148.65	-131.50
57	0.9390	262.78	1142.78	-128.01
58	0.9380	259.60	1141.60	-127.31
59	0.9376	259.60	1141.60	-127.31
60	0.9373	259.60	1141.60	-127.31

As a result of image matching, the image with the highest similarity (test image no.43 in this example) is set to p_2 , and whether it is outlier is determined. According to Chapter 3.2.3, the limit value of the pose change

during 30 frames of each site construction equipment is calculated as Table 4.5.

Table 4.5 Limits of pose change during 30 frames in train dataset (cm).

	Site 1	Site 2	Site 3
Arm_Bucket (Excavator)	234.65	259.96	316.00
Center (Dump Truck)	174.47	186.46	54.40

Since the value of Table 4.5 is the maximum value of the pose change during 30 frames, it is substituted based on the $p_2 - p_1$ frames according to Eq 3.

$$\begin{aligned}
 &(\text{pose of) no.43} - \text{no.20} = \\
 &\sqrt{(272.32 - 344.16)^2 + (1146.30 - 1161.09)^2 + (-130.1 + 142.57)^2} \\
 &= 74.40 < 234.65 \times \frac{43 - 20}{30} = 179.90 = 179.90
 \end{aligned}$$

Since p_2 is smaller than the limit value, it is not an outlier. Therefore, it is possible to interpolate the pose information between no.20 and no.43 using the values of p_1 and p_2 , and the interpolation result is shown in Table 4.6. Post-processing was performed in the same process for the pose estimation results of the remaining test datasets.

Table 4.6 Pose interpolation results from no.20 to no.43.

Image No.	Similarity	Arm_Bucket_x	Arm_Bucket_y	Arm_Bucket_z
20	0.9396	344.16	1161.09	-142.57
21	-	341.04	1160.45	-142.03
22	-	337.91	1159.80	-141.49
23	-	334.79	1159.16	-140.94
24	-	331.67	1158.52	-140.40
25	-	328.54	1157.87	-139.86
26	-	325.42	1157.23	-139.32
27	-	322.30	1156.59	-138.77
28	-	319.17	1155.95	-138.23
29	-	316.05	1155.30	-137.69
30	-	312.93	1154.65	-137.15
31	-	309.80	1154.02	-136.61
32	-	306.68	1153.37	-136.06
33	-	303.55	1152.73	-135.52
34	-	300.43	1152.09	-134.98
35	-	297.31	1151.44	-134.44
36	-	294.18	1150.80	-133.90
37	-	291.06	1150.16	-133.35
38	-	287.94	1149.52	-132.81
39	-	284.81	1148.87	-132.27
40	-	281.69	1148.23	-131.73
41	-	278.57	1147.59	-131.18
42	-	275.44	1146.94	-130.64
43	0.9446	272.32	1146.30	-130.10

4.3 Visualization

Visualization was performed for 2D and 3D, and 2D is a top-view representation of main keypoint's movement for construction equipment (e.g., excavator: arm_bucket, dump truck: center). However, excavator and dump truck are working in almost fixed positions in the collected data. Therefore, 2D visualization is described an example of dump truck leaving after the loading operation is finished. Figure 4.7 shows the movement of the dump truck from frame (a) to frame (b) as viewed from the top.

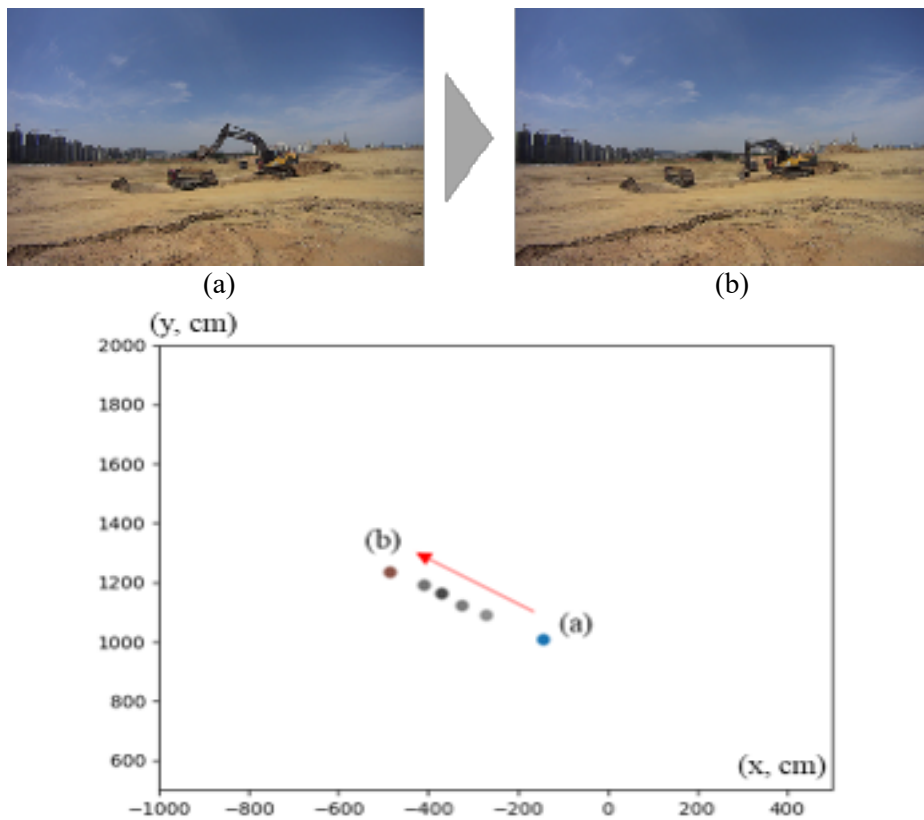
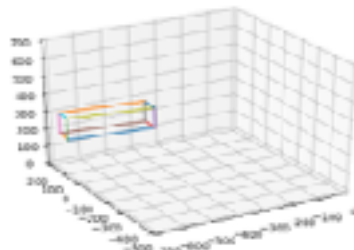
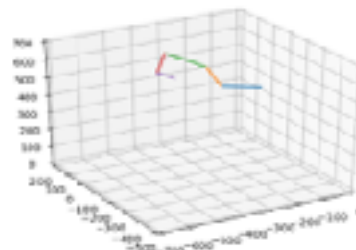


Figure 4.7 Example of 2D visualization (Site 1, dump truck).

3D visualization represents the 3D movement of the entire construction equipment by connecting keypoints of the equipment stored during annotation. The results of 3D visualizing the pose information calculated in Chapter 4.2 are as follows (Figure 4.8, 4.9, 4.10).

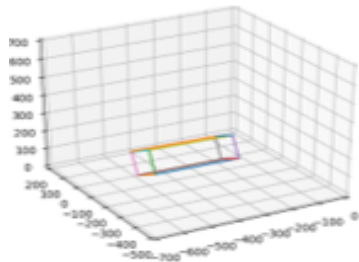


Dump truck

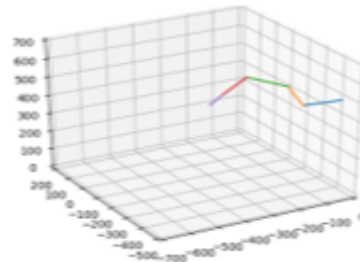


Excavator

Figure 4.8 Example of 3D visualization (Site 1).

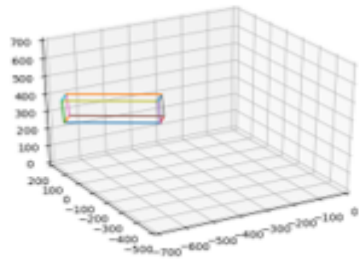


Dump truck

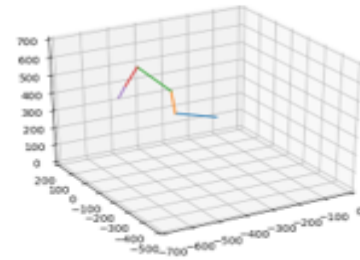


Excavator

Figure 4.9 Example of 3D visualization (Site 2).



Dump truck



Excavator

Figure 4.10 Example of 3D visualization (Site 3).

4.4 Evaluation of Validity

To evaluate validity of the experimental results, the root means square error (RMSE) equation was used. RMSE is a frequently used measure for validation in pose estimation. The RMSE equation is as below.

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^n (P_i - \hat{P}_i)^2 \quad \text{Eq.4}$$

n : number of keypoints

P_i : 3D (x, y, z) coordinates by keypoints in test images

\hat{P}_i : 3D (x, y, z) coordinates by keypoints in crop image data

n is the number of keypoints per construction equipment, in the case of an excavator, it has a value of 6 as Body, Center, Boom, Boom_Arm, Arm_Bucket, and Bucket, and in the case of a dump truck, it has a value of 8 as D_1 to D_8 (The center coordinates of the dump truck can be obtained by D_1 and D_8). P_i and \hat{P}_i are coordinates for each keypoint of the construction equipment 3D model saved in 30,000 images for each site through annotation using the Unity game engine in Chapter 3.1. P_i is the pose information of construction equipment for 9,000 test images that have performed pose estimation using an image matching algorithm, and \hat{P}_i is the pose information of construction equipment stored in 21,000 images used to build the crop image DB.

The value obtained by calculating the RMSE for each keypoint of construction equipment using the above equation Eq.4 for each site is shown in Table 4.7 and 4.8 below.

Table 4.7 RMSE for each keypoint (excavator).

	Site 1	Site 2	Site 3
Body	24.31	25.12	41.27
Center	31.23	37.35	35.49
Boom	33.44	41.73	52.15
Boom_Arm	43.28	47.54	68.15
Arm_Bucket	77.21	132.67	175.15
Bucket	74.36	112.45	154.82
Average	47.31	66.14	87.84

Table 4.8 RMSE for each keypoint (dump truck).

	Site 1	Site 2	Site 3
D_1	30.13	41.21	13.23
D_2	35.41	29.45	18.25
D_3	21.98	32.13	10.48
D_4	20.80	24.51	14.60
D_5	37.71	40.23	21.83
D_6	36.32	30.51	10.23
D_7	32.45	51.20	15.32
D_8	29.51	40.49	10.29
Average	30.54	36.22	14.28

4.5 Discussion

As a result of the experiment, it was confirmed that construction equipment (excavator, dump truck) was detected with fine performance in the test image and image matching. And visualizing the results of image matching can help identify where the current construction equipment is located at the construction site (i.e., 2D visualization), and the movement of the construction equipment at that location (i.e., 3D visualization). Furthermore the possibility of analyzing the interaction between an excavator and a dump truck in one virtual space was confirmed. Also, it showed good performance in the RMSE calculated for the evaluation of validity (Table 4.7 and 4.8). However, excavator's Arm_Bucket and Bucket parts showed relatively large RMSE values. It is judged that there is no image of the construction equipment having a similar posture to the test image in the crop image DB, or an error occurred due to an accuracy problem of annotation. In addition, not being completely overlaid because the model of the equipment used in the image and the model of the equipment used for the annotation are different, would have affected the accuracy of the annotation. Also, the model of the construction site image data and the 3D virtual model were not completely overlaid because they were different model and this would have affected the annotation accuracy.

Chapter 5. Conclusions

5.1 Summary and Contributions

In recent years, there is increasing practical interest in construction site monitoring using closed circuit television (CCTV) camera installed on site. So many researchers have developed various image analysis technologies for construction site monitoring. However, most of the previous studies did not cover the pose estimation of construction equipment. Knowing the pose information of construction equipment provides a more detailed view of the dynamic state and posture changes of the entire equipment, and it is possible to prevent potential collisions with nearby workers or equipment. There are limited studies to estimate the pose information of construction equipment, but these also have limitations (e.g., difficult to apply to the actual site). To overcome these problems, this research proposes a method of estimating and localizing the 3D pose of construction equipment using a 3D virtual model based on a single camera installed in the construction site.

This research has the following contributions. First, it proposed a method applicable to actual construction sites that solved the limitations of previous studies (e.g., many cameras need to be installed, IoT sensors should be installed for each driving part of construction equipment, etc.). Second, 3D pose information can be obtained without IoT sensors and for obscuring parts

of construction equipment. In previous studies that estimate the 3D poses of construction equipment through vision-based without IoT sensors, there is a problem that it is impossible to estimate the poses if the keypoints of obtaining the pose information are obscured or invisible. However, since the method proposed in this research uses a 3D virtual model to obtain a pose information using the shape of the entire equipment even if the corresponding keypoints are invisible, 3D pose estimation is possible. Finally, the dynamic state of the entire equipment can be understood by understanding the pose changes of the equipment, and pro-active action can be taken against potential safety accidents with the surrounding personnel and equipment.

5.2 Limitation and Future Study

In this research, a total of three sites were experimented, but each of them detected construction equipment using object detection models trained at the same site, and images were collected from one camera arrangement for each site. Image matching was performed within the crop image database built at the same site. If the trained model and the crop image database are experimented on a new site, the performance will be degraded. Also, if the resolution ratio of the camera used for annotation is different, the performance of pose estimation will be degraded.

In addition, there is a limitation that the 3D virtual model of construction equipment used in 2D - 3D annotation was different from the model of construction equipment in the collected construction site image data. The model of the excavator in the construction site image data used for the annotation was Volvo EC 300DL, and the 3D virtual model was the 3D Volvo EC 650 model. As a result, the 3D model was not fully matched to the construction equipment in the image data during the annotation (the same goes for dump trucks). When there is no image with a posture similar to the test image in the crop image database, there is a limitation that an error in image matching may occur. In addition, if there are multiple images having the same posture during annotation, the 3D model should be annotated so that all of the 3D models have the same pose information, but there is a limitation that a slight error occurs during annotation process.

As a future study, it could be to develop a methodology that can be

applied to a new site for an already established database. And it is possible to apply a deep learning method other than the template matching algorithm applied in this research. In addition, a method of constructing a crop image database by moving a 3D virtual model at various angles and projecting it as a 2D image, and performing image matching with the corresponding image and actual image data may be studied in the future.

Bibliography

- Azar, E., and McCabe, B. (2012). "Automated Visual Recognition of Dump Trucks in Construction Videos." *Journal of Computing in Civil Engineering*, 26(6), pp. 769-781.
- Azar, E., Dickinson, S., and McCabe, B. (2013). "Server-Customer Interaction Tracker: Computer Vision-Based System to Estimate Dirt-Loading Cycles." *Journal of Construction Engineering and Management*, 139(7), pp. 785-794.
- Azar, E., Feng, C., and Kamat, V. (2015). "Feasibility of in-plane articulation monitoring of excavator arm using planar marker tracking" *Journal of Information Technology in Construction*, 20, pp. 213-229.
- Bugler, M., Borrmann, A., Ogunmakin, G., Vela, P., and Teizer, J. (2016). "Fusion of Photogrammetry and Video Analysis for Productivity Assessment of Earthwork Processes." *Computer-Aided Civil and Infrastructure Engineering*, 32, pp. 107-123.
- Chen, C., Zhu, Z., and Hammad, A. (2020). "Automated excavators activity recognition and productivity analysis from construction site surveillance videos." *Automation in Construction*, 110, 103045.
- Choi, N., Son, H., Kim, C., Kim, C., and Kim, H. (2008). "Rapid 3D object recognition for automatic project progress monitoring using a stereo vision system." *The 25th International Symposium on Automation and*

Robotics in Construction, pp. 58-63.

Deng, H., Hong, H., Luo, D., Deng, Y., and Su, C. (2019) “Automatic Indoor Construction Progress Monitoring for Tiles Based on BIM and Computer Vision.” *Journal of Construction Engineering and Management*, 146(1), 04019095.

Dimitrov, A., and Golparvar-Fard, M. (2014). “Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections.” *Advanced Engineering Informatics*, 28, pp. 38-49.

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

Fang, Y., Cho, Y., and Chen, J. (2016). “A framework for real-time pro-active safety assistance for mobile crane lifting operations.” *Automation in Construction*, 72(3), pp. 367-379.

Guo, Y., Niu, H., and Li, S. (2018). “Safety Monitoring in Construction Site based on Unmanned Aerial Vehicle Platform with Computer Vision using Transfer Learning Techniques.” *7th Asia-Pacific Workshop on Structural Health Monitoring*, pp. 1052-1060.

Jiang, Z., and Messner, J. (2020). “Computer Vision-Based Methods Applied to Construction Processes: A Literature Review” *Construction Research Congress 2020*, pp. 1233-1241.

- Kang, J., Choi, P., and Eom, D. (2018). "Precise Bucket Pin-point Estimation of Excavator in 3-dimensional by Integration of Accelerometers with RTK GPSs." *Institute of Control, Robotics and Systems*, 24(10), pp. 930-938.
- Kim, H., Bang, S., Jeong, H., Ham, Y., and Kim, H. (2018). "Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation." *Automation in Construction*, 92, pp. 188-198.
- Kim, H., Kim, H., Hong, Y., and Byun, H. (2017). "Detecting Construction Equipment Using a Region-Based Fully Connected Network and Transfer Learning." *Journal of Computer in Civil Engineering*, 32(2), 04017082.
- Kim, H., Kim, K., and Kim, H. (2016). "Vision-Based Object-Centric Safety Assessment Using Fuzzy Interference: Monitoring Stuck-By Accidents with Moving Objects." *Journal of Computer in Civil Engineering*, 30(4), 04015075.
- Kim, J., and Chi, S. (2020). "Multi-camera vision-based productivity monitoring of earthmoving operations." *Automation in Construction*, 112, 103121.
- Kim, J., Chi, S., and Seo, J. (2018). "Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks." *Automation in Construction*, 87, pp. 297-308.
- Lee, J., Kim, B., Sun, D., Han, C., and Ahn, Y. (2019) "Development of

Unmanned Excavator Vehicle System for Performing Dangerous Construction Work” *Sensors*, 19(22), 4853.

Li, H., Lu, M., Hsu, S., Gray, M., and Huang, T. (2015). “Proactivebehaviour-based safety management for construction safety improvement.” *Safety Science*, 75, pp. 107-117.

Liang, C., Lundeen, K., McGee, W., Menassa, C., Lee, S., and Kamat, V. (2019). “A vision-based marker-less pose estimation system for articulated construction robots.” *Automation in Construction*, 104, pp. 80-94.

Liang, C., Kamat, V., and Menassa, C. (2018) “Real-Time Coonstruction Site Layout and Equipment Monitoring” *Construction Research Congress 2018*, pp. 64-74.

Lundeen, K., Dong, S., Fredricks, N., Akula, M., Seo, J., and Kamat, V. (2016). “Optical marker-based end effector pose estimation for articulated excavators.” *Automation in Construction*, 65, pp. 51-64

Luo, H., Li, H., Cao, D., Dai, F., Seo, J., and Lee, S. (2018). “Recognizing Diverese Construction Activities in Site Images via Relevance Networks of Construction-Related Objects Detected by Convolutional Neural Networks.” *Journal of Computer in Civil Engineering*, 32(3), 04018012.

Luo, H., Wang, M., Wong, P., and Cheng, J. (2020). “Full body pose estimation of construction equipment using computer vision and deep

- learning techniques.” *Automation in Construction*, 110, 103016.
- Luo, H., Wang, M., Wong, P., Tang, J., and Cheng, J. (2020). “Construction machine pose prediction considering historical motions and activity attributes using gated recurrent unit (GRU)” *Automation in Construction*, 121, 103444.
- Pentek, Z., Hiller, T., Liewald, T., Kuhlmann, B., and Czmerk, A. (2017). “IMU-based mounting parameter estimation on construction vehicles.” *2017 DGON Inertial Sensors and Systems(ISS)*, pp. 1-14.
- Pradhananga, N., and Teizer, J. (2012) “GPS-based framework towards more realistic and real-time construction equipment operation simulation.” *In Proceedings of the 2012 Winter Simulation Conference*, 64.
- Rashid, K., and Louis, J. (2019) “Construction Equipment Activity Recognition from IMUs Mounted on Articulated Implements and Supervised Classification.” *International Conference in Computing in Civil Engineering 2019*
- Roberts, D., and Golparvar-Fard, M. (2019) “End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level” *Automation in Construction*, 105, 102811.
- Seo, J., Han, S., Lee, S., and Kim, H. (2015). “Computer vision techniques for construction safety and health monitoring.” *Advanced Engineering Informatics*, 29, pp. 239-251.

- Soltani, M., Zhu, Z., and Hammad, A. (2016). "Towards Part-Based Construction Equipment Pose Estimation Using Synthetic Images." Construction Research Congress 2016, pp. 980-989.
- Soltani, M., Zhu, Z., and Hammad, A. (2017). "Skeleton estimation of excavator by detecting its parts" Automation in Construction, 82, pp. 1-15.
- Soltani, M., Zhu, Z., and Hammad, A. (2018). "Framework for location data fusion and pose estimation of excavators using stereo vision." Journal of Computing in Civil Engineering, 32(6), 04018045.
- Souma-Gyimah, G., Frimpong, S., Nyaaba, W., and Gbadam, E. (2019) "A computer vision system for terrain recognition and object detection tasks in mining and construction environments" SME Annual Conference.
- Sun, D., Kim, S., Lee, Y., Lee, S., and Han, C. (2017) "Pose and Position Estimation of Dozer Blade in 3-dimensional by Integration of IMU with Two RTK GPSs" 34th International Symposium on Automation and Robotics in Construction (ISARC 2017).
- Tang, J., Luo, H., Wong, P., and Cheng, J. (2020). "Study of IMU Installation Position for Posture Estimation of Excavators." International Conference on Computing in Civil and Building Engineering, 98, pp. 980-991.
- Teizer, J., Lao, D., and Sofer, M. (2007) "Rapid automated monitoring of

construction site activities using ultra-wideband” 24th International Symposium on Automation & Robotics in Construction (ISARC 2007), pp. 23-28.

Tuttas, S., Braun, A., Borrmann, A., and Stilla, U. (2016) “Evaluation of acquisition strategies for image-based construction site monitoring” The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 40(3), 321.

Vahdatikhaki, F., and Hammad, A. (2015). “Dynamic equipment workspace generation for improving earthwork safety using real-time location system” *Advanced Engineering Informatics* 29, pp. 459-471.

Vahdatikhaki, F., Hammad, A., and Siddiqui, H. (2015). “Optimization-based excavator pose estimation using real-time location systems.” *Automation in Construction*, 56, pp. 76-92.

Wang, Z., Zhang, Q., Yang, B., Wu, T., Lei, K., Zhang, B., and Fang, T. (2020). “Vision-Based Framework for Automatic Progress Monitoring of Precast Walls by Using Surveillance Videos during the Construction Phase.” *Journal of Computing in Civil Engineering*, 35(1), 04020056.

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

Zhang, C., Hammad, A., and Rodriguez, S. (2012). “Crane Pose Estimation Using UWB Real-Time Location System.” *Journal of Computing in*

Civil Engineering, 21(6), pp. 625-637.

초 록

최근 영상분석기술의 빠른 발전과 함께 현장에 설치된 closed circuit television (CCTV) 카메라를 활용하는 건설현장 영상 모니터링 기술에 대한 실무적 관심이 높아지고 있다. CCTV 영상 데이터를 활용한 건설장비의 작업 분류 및 생산성 분석, 위험구역 접근 감지 등 건설현장 영상 모니터링을 위한 다양한 영상분석기술들도 개발되었다. 하지만 이전의 연구들은 주로 영상 데이터 속에서 건설장비를 식별하고 추적하거나, 작업을 분류하는 데 초점을 맞추었을 뿐 건설장비의 포즈 추정에 대해서는 자세히 다루지 않았다.

건설장비의 포즈 추정은 장비의 각 키포인트별 2차원 혹은 3차원 좌표 정보 (i.e., its location and orientation)를 획득하는 것으로, 건설장비 자세 파악 및 원격 제어, 건설 현장의 안전성 분석 및 건설 프로젝트의 생산성 분석을 위한 기본적인 기계 정보를 제공한다. 건설장비의 포즈 정보를 알면 장비 전체의 동적 상태 및 자세 변화를 보다 자세하게 파악할 수 있으며, 주변 작업자나 장비와의 잠재적 충돌 사고를 방지하는 것이 가능하다.

본 연구는 단일 카메라 영상을 기반으로 3차원 가상 모델을 이용하여 건설장비의 3차원 포즈와 위치를 추정하는 방법을 제안한다. 연구는 크게 네 가지 단계로 구성되어 있다. 첫째로,

문헌조사를 통해 기존 영상분석기술을 활용한 건설현장 모니터링 기술과 건설장비 포즈 추정 기술을 정의하였다. 둘째로, 단일 카메라 영상에서 건설장비의 포즈를 추정하기 위해 건설현장 2차원 이미지 데이터에 3차원 가상 모델 데이터를 저장하였다. 셋째로, 건설현장 영상 데이터로부터 건설장비를 탐지 및 추출하고, 이미지 매칭을 통하여 유사 이미지를 도출하는 건설장비 포즈 추정 방법을 개발하였다. 마지막으로, 3곳의 건설현장에서 수집한 영상 데이터를 이용하여 건설장비 포즈 추정 결과의 성능을 검증하였다.

그 결과 본 연구에서 제안한 방법을 통해 단일 영상 데이터에서 건설장비의 포즈 추정할 수 있음을 보였으며, 건설장비가 가려지는 부분에 대해서도 포즈 추정이 가능함을 확인하였다. 뿐만 아니라, 여러 장비들의 포즈 추정값을 이용하여 장비들 간의 상호작용 분석에도 도움이 될 것으로 기대된다.

주요어: 단일 카메라 기반, 3차원 가상 모델, 3차원 포즈 추정, 건설장비

학 번: 2019-27248