sensors

MDPI

Article

Pose Estimation of Excavator Manipulator Based on Monocular Vision Marker System

Jiangying Zhao, Yongbiao Hu * and Mingrui Tian

National Engineering Laboratory for Highway Maintenance Equipment, Chang'an University, Xi'an 710064, China; jyzhao.chd@chd.edu.cn (J.Z.); tianmingrui1020@163.com (M.T.)
* Correspondence: hybiao@chd.edu.cn

Abstract: Excavation is one of the broadest activities in the construction industry, often affected by safety and productivity. To address these problems, it is necessary for construction sites to automatically monitor the poses of excavator manipulators in real time. Based on computer vision (CV) technology, an approach, through a monocular camera and marker, was proposed to estimate the pose parameters (including orientation and position) of the excavator manipulator. To simulate the pose estimation process, a measurement system was established with a common camera and marker. Through comprehensive experiments and error analysis, this approach showed that the maximum detectable depth of the system is greater than 11 m, the orientation error is less than 8.5°, and the position error is less than 22 mm. A prototype of the system that proved the feasibility of the proposed method was tested. Furthermore, this study provides an alternative CV technology for monitoring construction machines.

Keywords: excavator pose estimation; monocular camera; marker; error analysis; computer vision technology

Citation: Zhao, J.; Hu, Y.; Tian, M. Pose Estimation of Excavator Manipulator Based on Monocular Vision Marker System. *Sensors* **2021**, *21*, 4478. https://doi.org/10.3390/s21134478

Academic Editors: Michał R. Nowicki, Giorgio Grisetti and Marco Camurri

Received: 12 May 2021
Accepted: 25 June 2021
Published: 30 June 2021

1. Introduction

The construction industry has, for a long time, faced issues surrounding low productivity [1] and high hazard ratios. According to the Ministry of Housing and Urban-Rural Development of the People's Republic of China [2], 8.41% of accidents were caused by construction machines in China's construction industry (in 2019). Therefore, it is important to monitor machinery for the management of construction production.

To address this challenge, construction automation (CA) [3] technology has been looked at as a means to improve productivity and safety in construction, which could mitigate dependence on skilled workers, enable safer collaborations between construction machinery and surrounding workers, reduce accidents, and improve productivity. CA has been gradually applied in the construction industry, i.e., by monitoring the environment of construction sites and tracking the movement of construction machines.

On construction sites, CA (e.g., sensor technology) could be used to monitor the construction site environment in real time and return detected information to the operator, which would help the operator understand the surrounding environment (e.g., construction workers and workplace) better. In this way, construction could be conducted more safely and efficiently. On the other hand, monitoring the movement of construction machines (i.e., the pose of an excavator, which includes orientation), combined with other technologies (e.g., task decision and motion planning technology, target detection, and tracking technology), could make construction machines perform actions (e.g., send a warning sound or machine emergency braking) independently, based on the surrounding environment.

There are two primary ways of monitoring construction machines: vision-based [4,5] and sensor-based (i.e., not vision-based) methods. Compared to sensor-based methods, vision-based methods (e.g., UWB, GPS, IMU) [6–8] have economic advantages (i.e., they

are affordable for many construction contractors to implement). In addition, a vision-based method could provide operators with rich information of the construction sites (e.g., interaction between workers and entities, and wide visual range), which could help operators complete construction projects quickly and accurately. As excavators are the most used construction machines, and their poses are constantly changing, they threaten the safety of workers and surrounding entities. Thus, the precise pose of a construction machine could provide reference information for the operator to adjust excavation work in real time. Furthermore, it could potentially be applied in construction management (such as safety monitoring and productivity analysis).

Therefore, it is necessary to monitor the pose of the excavator manipulator in real time, from the perspective of work efficiency and construction safety. In this paper, the approach, based on a monocular vision and marker, was proposed to estimate the excavator manipulator pose. A prototype was established to model the measurement system, which presents sufficient accuracy for implementation through error analysis. As shown in Figure 1, this system mainly consists of a monocular camera, marker, and image processing system.

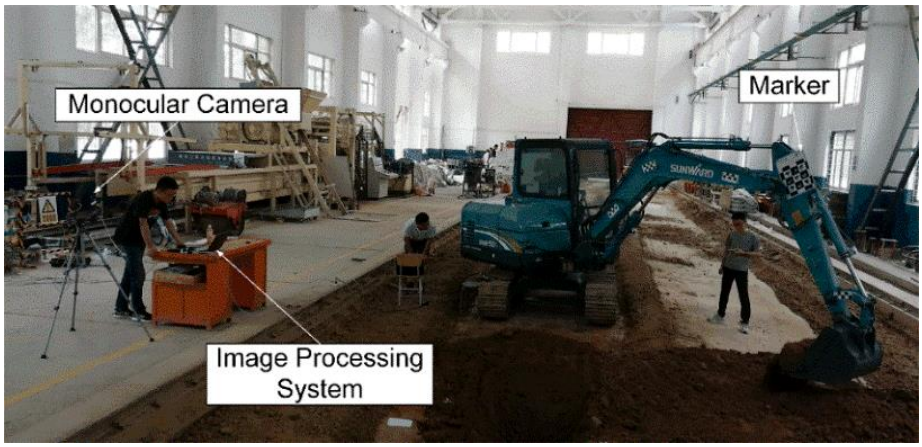


Figure 1. Excavator pose estimation system.

This work presents the following contributions: first, the authors propose a method of estimating the excavator pose based on a monocular marker system; second, our method could accurately estimate 6D poses rather than a 2D image position, which provides more comprehensive information for construction applications; finally, in this paper, we present a prototype and analyze the accuracy of the pose estimation of this method, showing its feasibility in practical applications.

The remaining parts of this paper are organized as follows: Section 2 reviews work related to the vision-based pose estimation of the excavator; Section 3 presents the pose estimation approach; Section 4 discusses the details of the measurement system design based on the proposed pose estimation approach; Section 5 presents the experimental results and discussion. Finally, the author's conclusions and future work are summarized in Section 6.

2. Related Work

In previous studies, CV technology, such as object detection, object tracking, and pose estimation, was often applied to monitor construction equipment. A complete review of these works is beyond the scope of this paper. Instead, more related studies (e.g., excavator pose estimation) are reviewed in this section. In current practices, there are two types of pose estimation methods used in construction machinery: marker-less and marker-based.

Marker-less methods only require image acquisition equipment and a processing system, which tracks the excavator by extracting image features. Zou and Kim [9] proposed a system to quantify the idle time of hydraulic excavators through the hue, saturation, value (HSV) color space, which could separate excavators from the background. At the

same time, the system did not estimate the object's movement parameters. Rezazadeh Azar and McCabe [10] used a classifier based on histogram of oriented gradient (HOG) detectors to identify excavators with six poses. Then, the server-customer interaction tracker (SCIT) [11] was developed. This system not only identifies the relative interaction between the excavator and truck, but could also measure excavator loading cycles. However, it fails to provide an accurate position and orientation parameters. Yuan et al. [12,13] proposed a binocular vision framework based on a hybrid kinematic shape and key node features to detect, track, and locate the excavators. As for the visual range and depth errors of the binocular vision, it is heavily dependent on a fixed baseline distance. In addition, this system detects key nodes covered by multi-view templates. Thus, the detection system has low robustness, due to the different templates generated at different angles. Soltani et al. [14] presented a method to detect the parts of the excavator manipulator using synthetic images to train the detectors, and then the 2D skeleton of the excavator was extracted. This method requires a synthetic dataset to train the detector and only estimates the 2D orientation [15]. Xu et al. [16,17] described a neural network-based structure to estimate excavator cylinder displacements. Simulation results illustrate that the system could be used to measure the manipulator position, but neural network training requires a lot of data and lacks orientation estimation. Liang et al. [18] proposed a marker-less pose estimation system for 2D and 3D construction robots, which uses the Stacked Hourglass Network [19] of the human pose estimation method. Although this system could detect the robot pose, as well as other methods, it needs to collect a lot of data, and the accuracy of the 3D pose estimation is low [20]. Therefore, further optimization of the algorithm is needed for excavation applications. Torres Calderon et al. [21] used a method of generating excavator pose data through the virtual environment of the Unity 3D engine. The generated dataset is used to train a network model based on deep learning, while analyzing only excavator activity and action prediction. Fang et al. [22] developed a monocular vision technology that combines semantics and prior knowledge to locate architectural entities, using deep learning algorithms to segment the instances in the image, and then foreknowledge models to estimate location information. However, this method requires a large amount of labeled datasets, and position accuracy is limited. Luo et al. [23,24] trained three convolutional neural networks (CNNs) based on the dataset labeled with 2D position information, which was used to estimate the pose of the construction equipment by detecting the key points of the excavator. Subsequently, an recurrent neural network (RNN) named gated recurrent unit (GRU) was proposed, which combined historical data of device movement and activity attributes for pose prediction [25]. However, the above methods need a large amount of data with labeled pose information, which is time consuming and labor intensive. Moreover, as the pose error is relatively large, it is difficult to provide a satisfactory 6D pose.

Marker-based methods detect the marker installed on the excavator manipulator, and the manipulator pose is obtained through CV technology. Compared to marker-less methods, marker-based methods could commonly achieve greater accuracy of pose estimation. Rezazadeh Azar et al. [26] presented a configuration based on a marker-based detecting algorithm to measure the pose of the excavator manipulator [27]. However, this configuration worked with the camera's optical axis, to be "normal" to the marker plane. Lundeen et al. [28,29] developed an excavator pose measurement system that consists of a marker and an optical system. The system designed four solutions to estimate the final effector pose of the excavator. Therefore, the implementation of the system at construction sites is difficult due to the complexity of the electromechanical system. Feng et al. [30] proposed a structure of camera marker networks. This method shows the feasibility of the system, and the accuracy of the centimeter level measurement was achieved [31]. However, the occlusion resistance was weak, and the orientation error analysis was not considered.

Previous works have shown promising results in excavator monitoring. However, there are still some limitations, such as short visual range or heavy maintenance of binocular cameras and RGB-D cameras. Moreover, few of them focus on accurate monitoring of

the overall pose (i.e., position and orientation). Compared to other vision-based methods, the monocular camera is flexible to deploy and simple to maintain on construction sites. Furthermore, the marker helps to increase environment information compared to marker-less methods. Therefore, this work proposes a measurement system that consists of a monocular vision marker system based on the pose estimation method.

3. Pose Estimation Approach

Before presenting the structure of the measurement system, the approach of estimating the pose based on the marker is explained. The overview of the approach is shown in Figure 2, which is composed of camera calibration and pose estimation.

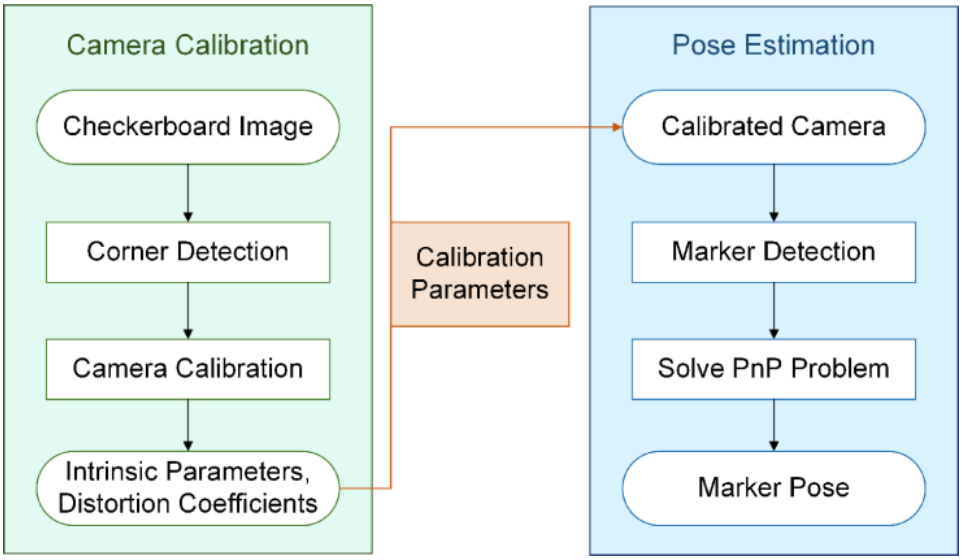


Figure 2. Overview of the pose estimation approach.

3.1. Camera Calibration

A camera should be calibrated prior to starting the measurement system to estimate the excavator pose, as shown in the left half of Figure 2. Camera calibration is a critical process to improve system measurement accuracy.

Camera calibration aims to obtain intrinsic parameters, including camera focal lengths in the x and y directions f_x, f_y, and the main point (the intersection of the image coordinate system plane and the optical axis) (u_0, v_0), and distortion coefficients k_1, k_2, k_3, p_1, p_2. This is achieved in the MATLAB Camera Calibrator Toolbox based on Zhang's calibration method [32]. It is important to note that, although the pose could also be optimized together with the extrinsic parameters (rotation matrix and translation vector), the optimization solution gets worse with the increase of unknown variables. Therefore, in this paper, intrinsic parameters and distortion coefficients are calibrated before pose estimation.

3.2. Pose Estimation

There are several typical marker-based pose estimation methods, including ARTag [33], AprilTag [27], and CALTag [34]. CALTag is a self-identifying marker that could be accurately detected in the image. Compared to other markers, CALTag demonstrated better resistance to occlusion (e.g., CALTag size of 98 × 98 mm had an 88% recognition rate with 25.5% to 32.5% occlusion) [35].

As shown in the right half of Figure 2, the pose is estimated by solving the Perspective-n-Points (PnP) problem [36–40]; that is, estimating the relative orientation and position information between the 3D points in the world and points of the corresponding 2D image.

The solution to the PnP problem is based on the camera model (as shown in Figure 3), which is expressed as the following Equation (1):

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix}, \tag{1}$$

where $\begin{bmatrix} u & v \end{bmatrix}^{\mathrm{T}}$ are pixel coordinates, $\begin{bmatrix} X_W & Y_W & Z_W \end{bmatrix}^{\mathrm{T}}$ are world coordinates, s is called the scale factor, $\boldsymbol{R}$ and $\boldsymbol{t}$ are the rotation matrix and the translation vector, respectively.

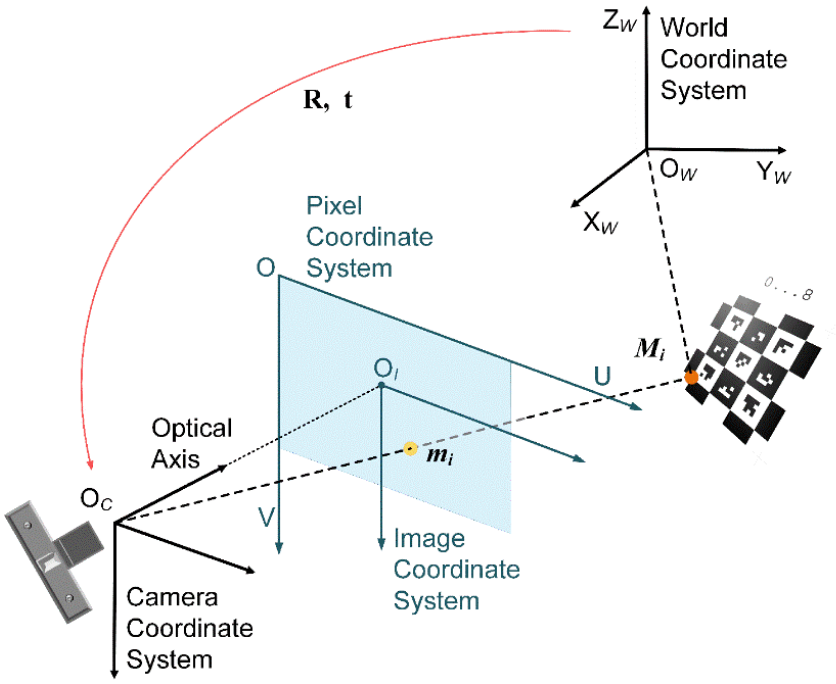


Figure 3. Camera model.

Equation (1) can be written as:

$$s\boldsymbol{p} = \boldsymbol{K}\begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \end{bmatrix}\boldsymbol{P}, \tag{2}$$

where $\boldsymbol{K}$ is called the camera's intrinsic matrix (usually calibrated before optimization), $\boldsymbol{P}$ denotes the 3D point, and $\boldsymbol{p}$ denotes the corresponding 2D point.

The pose is solved by minimizing the re-projection error, which is, minimizing the following function:

$$\underset{\boldsymbol{R}_i, \boldsymbol{t}_i}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \hat{\boldsymbol{p}}\left(\boldsymbol{K}, \boldsymbol{R}_i, \boldsymbol{t}_i, \boldsymbol{P}_j\right) - \boldsymbol{p}_{ij} \right\|^2, \tag{3}$$

where $\hat{\boldsymbol{p}}\left(\boldsymbol{K}, \boldsymbol{R}_i, \boldsymbol{t}_i, \boldsymbol{P}_j\right)$ is the projection of the point $\boldsymbol{P}_j$ in the i-th image.

Solving this nonlinear least squares problem is also called bundle adjustment [40]. The Levenberg–Marquardt algorithm is generally used to solve this problem and the g2o solver [41] could be adopted during the solution of this problem. It should be noted that the EPnP [38] method is used to estimate the camera pose as the initial solution, and then bundle adjustment is used to optimize the estimated value.

3.3. Error Analysis

It is inadequate to just estimate pose in a monocular vision marker system. The accuracy of the system needs to be evaluated, which could provide a reference to guide

practice. In this paper, the accuracy of the system is evaluated by calculating the absolute error. Given the true camera's true pose and the corresponding pose estimate value, the absolute orientation and position errors are calculated by the following equations:

$$E_{\text{rot}} = \left| \boldsymbol{r}^{l}_{\text{true}} - \boldsymbol{r}^{l} \right|, \tag{4}$$

$$E_{\text{trans}} = \left| \boldsymbol{t}_{\text{true}} - \boldsymbol{t} \right|, \tag{5}$$

where $\boldsymbol{r}^{l}_{\text{true}}$ and $\boldsymbol{r}^{l}$ are the ground truth and the estimated value of the rotation matrix in the form of Euler angle (roll–pitch–yaw angle), and l takes roll, pitch and yaw angles, respectively. Likewise, $\boldsymbol{t}_{\text{true}}$ and $\boldsymbol{t}$ are the ground truth and the estimated value of the translation vector.

4. Design of the Measurement System

Based on the theory described in the previous section, the measurement system was designed with a monocular camera and marker, and in this section, the details of this system will be presented.

4.1. Marker Layout

To obtain the excavator manipulator pose, the marker needs to be installed on the manipulator. First, it needs to determine the size of the CALTag. The size of the CALTag should fit the size of the excavator's manipulator. It is not convenient to install a larger CALTag on the manipulator. In addition, a smaller CALTag could cause the measurement system to track failure in a long-distance operation. Second, the CALTag should be installed flat on the manipulator, which is useful for the camera to accurately capture the marker. Finally, the marker layout is determined, ensuring that the marker is within the camera's field of view based on multiple excavation operations.

4.2. Monocular Vision Marker System Design

There are two ways to estimate the pose using the camera to capture the marker. One way is to install the camera on a tripod and then attach it to the side of the excavator (Figure 4). The placement of the monocular vision system must ensure that the marker is always in the camera's field of view during the tracking process of the measurement system. As long as the camera is attached to this system and the marker could be detected in real time, the marker pose could be estimated. The other way is to attach a camera to the excavator cab (Figure 5), to ensure the marker is always within the camera's field of view, meeting the real-time tracking requirements of the manipulator.

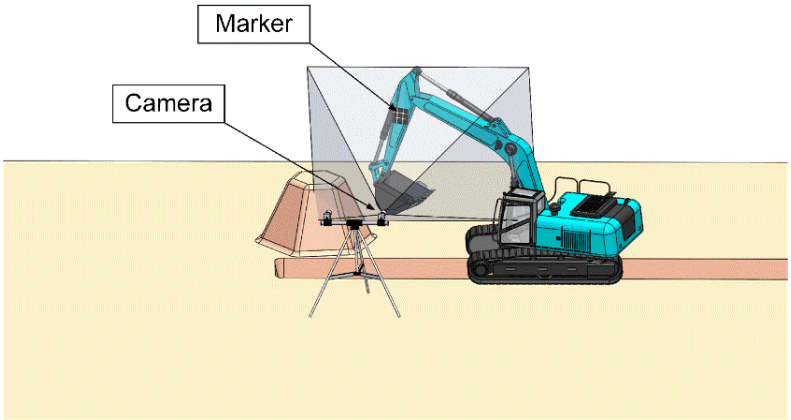


Figure 4. The camera is attached to the side of the excavator.

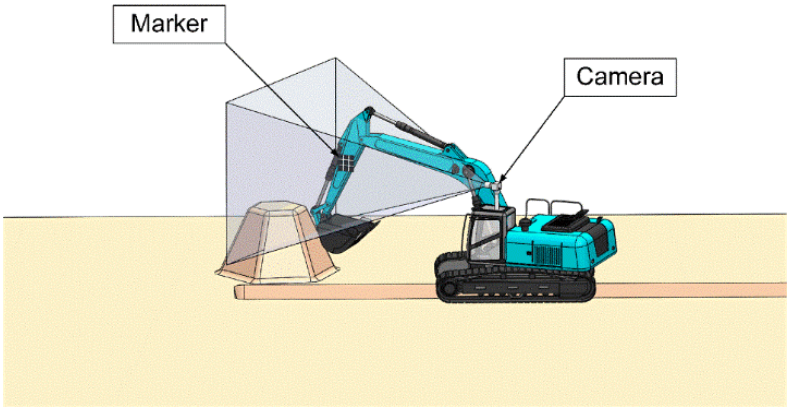


Figure 5. The camera is attached to the excavator cab.

However, in practice, certain operations (e.g., land leveling and long-distance trenching) would cause the marker to be out of the camera's field of view when the camera was placed on the side of the excavator, and the measurement system could not track the marker, causing the system to crash. For this problem, a possible solution is to install more cameras (i.e., single marker and multiple camera system) taking into account the construction site environment. Figures 4 and 5 only illustrate the simple design of the system. All of this leads to system design.

4.3. System Prototype

To verify the feasibility of the monocular vision marker system (Figure 1), the experimental model of the measurement system is established. In this paper: first, the marker is installed on a rotatable plate to simulate the movement of the excavator; second, a camera installed on a tripod is adopted to track the marker in real time; finally, the excavator pose is calculated by the image processing system.

5. Results and Discussion

In this section, three sets of experiments were performed and the results were analyzed. In addition, the authors discuss limitations and applicability compared to existing vision-based pose estimation methods.

5.1. Results

5.1.1. Effectiveness Experiments of Measurement System

Before implementing this system to estimate marker pose, a group of experiments was carried out under different conditions to assess the effectiveness of the measurement system.

Considering the size of the excavator manipulator (a 5-ton excavator, with a manipulator size of 200 to 400 mm in width), a CALTag with a size of 228 × 228 mm was adopted as the marker, and the image resolution was 1280 × 960 pixels. It should be noted that the normal lighting condition was considered only in these experiments. As described in the previous section, the camera should be calibrated before estimating the pose. Table 1 lists the calibration results.

Table 1. Results of the calibration of intrinsic parameters and distortion coefficients.

Calibration Parameters		Calibration Results
Intrinsic parameters (pixels)	(f_x, f_y)	(3172.2, 3144.8)
	(u_0, v_0)	(590.0959, 485.6238)
Distortion coefficients	(k_1, k_2, k_3)	(0.0353, −0.7045, −0.7541)
	(p_1, p_2)	(−0.0015, −0.0043)

Then, the effectiveness of the system was evaluated by calculating the maximum detectable depth. Since distance and angle are the vital factors affecting effectiveness, in this paper, the experiments were performed while the CALTag pitch angle was 0° and 45°, respectively. The experimental results are shown in Figure 6, the maximum detectable depth is greater than 11 m with the 0° pitch (Figure 6a). Furthermore, it could be seen that, when approaching the maximum depth, although the marker could be detected, the absolute error between the ground truth and the estimated depth changes abruptly; this point is considered an outlier (i.e., the maximum detectable depth). Likewise, when the pitch of the marker is 45° (Figure 6b), the maximum detectable depth is about 8 m, which is significantly less than the result of the pitch of the CALTag being 0°.

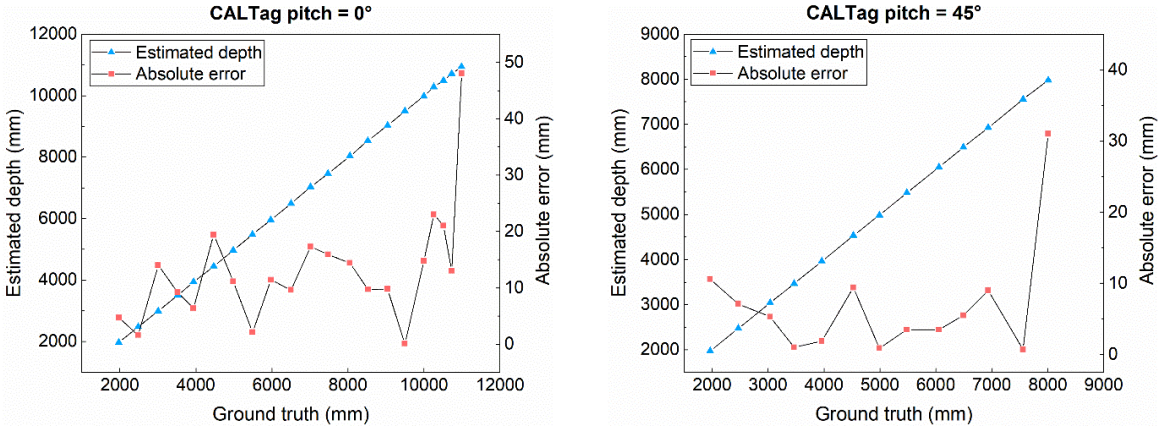


Figure 6. Maximum detectable depth under pitch difference. (**a**) Estimated depth under CALTag pitch = 0°. (**b**) Estimated depth under CALTag pitch = 45°.

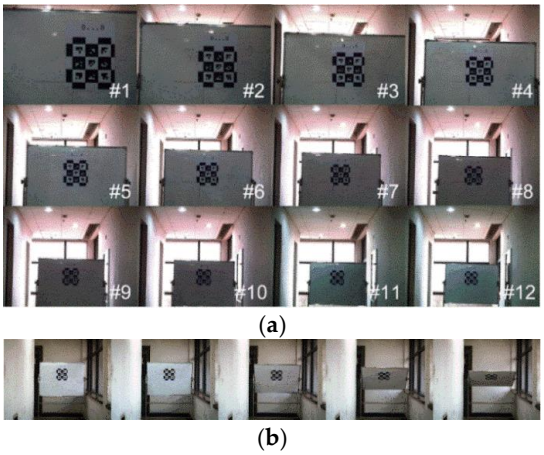
5.1.2. Orientation Accuracy Experiments

In this paper, we evaluate accuracy by the orientation error that is calculated using Equation (4). When measuring the measurement's accuracy, the camera was fixed, and mounted at the same height as the bottom-left of the marker. The marker plane was adjusted to be perpendicular to the normal direction of the camera lens, which was free to rotate about a horizontal axis. The orientation angle was measured by the multifunctional angle ruler (with the accuracy of 1/30°) as the ground truth when the pitch angle varied, which compared with the estimated value using the pose estimation approach. These experiments are shown in Figure 7, which shows marker detection under different depth and pitch conditions.

Figure 8 shows the absolute error of orientation in different configurations, in these two box plots, the maximum absolute pitch error is less than 8.5° and the average absolute pitch error is about 4° in different configurations.

5.1.3. Position Accuracy Experiments

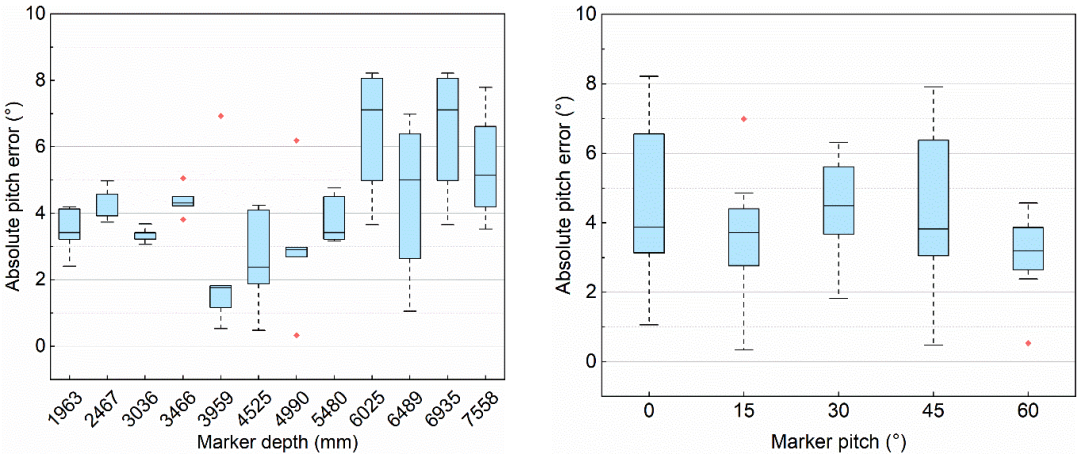
In this paper, vertical displacement (i.e., depth) was selected because ground truth measurements could be quickly and accurately obtained using a laser rangefinder, and because depth (z-axis) is generally the component of greatest interest in most excavation activities (and a critical factor affecting the method's feasibility). Position accuracy was completed by calculating the absolute depth error through Equation (5). The estimated value of the measurement system was compared with the ground truth, which was measured by laser rangefinder (with an accuracy of 1 mm). These experiments were tested in different configurations. Figure 9 shows the marker detection under different pitch and depth conditions.



(**a**)

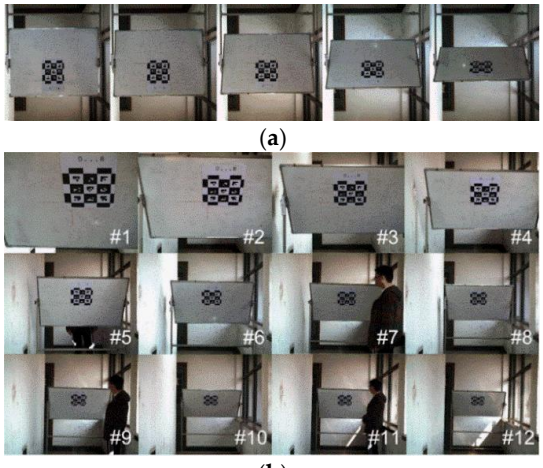
(**b**)

Figure 7. Orientation accuracy experiments with different configurations. (**a**) Different depth. (**b**) Different pitch. # 1 to 12 represents change in depth from 1963 to 7558 mm.



(**a**) (**b**)

Figure 8. The absolute pitch error with different configurations. (**a**) Absolute pitch error under different marker depths. (**b**) Absolute pitch error under different marker pitches.



(**a**)

(**b**)

Figure 9. Position accuracy experiments with different configurations. (**a**) Different marker pitch. (**b**) Different marker depth. # 1 to 12 represents change in depth from 1963 to 7558 mm.

As shown in Figure 10, the maximum absolute depth error is less than 22 mm and the average absolute depth error of the pitch angle is about 7 mm. Another prototype experiment (as shown in Figure 1) shows that the absolute maximum depth error is less than 25 mm, which could meet the construction requirements.

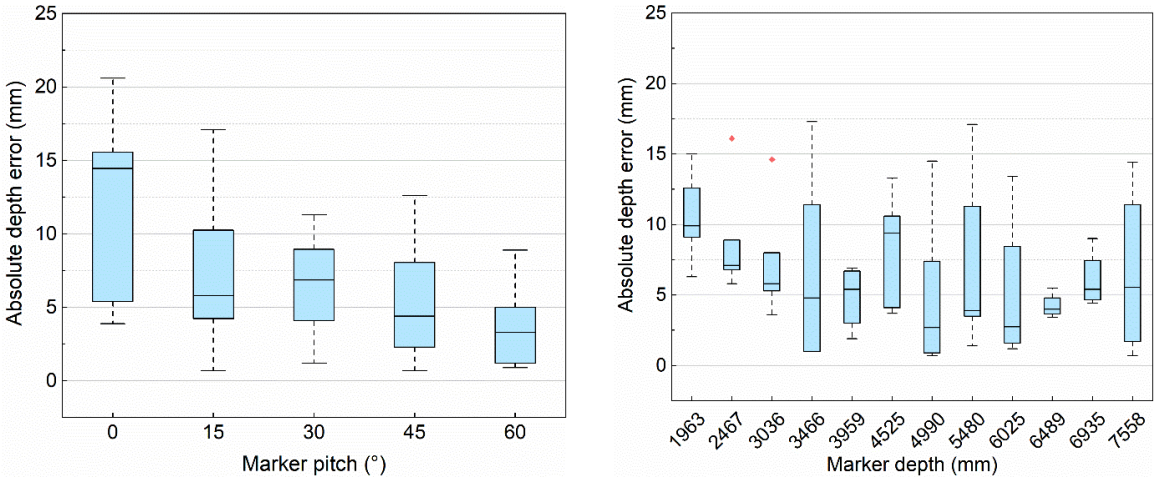


Figure 10. The absolute depth error with different configurations. (**a**) Absolute depth error under different marker pitches. (**b**) Absolute depth error under different marker depths.

5.2. Discussion

According to the experimental results, it could be verified that the pose estimation approach is viable in the monitoring of construction machines. While the maximum detectable depth of this work (the error is less than 50 mm when depth is about 11 m) is less than the previous work (the error was 0.5 m when the depth was 15 m), from the results of Yuan's work [12,13]. This is because marker-based methods allow for better ability to capture key nodes compared to the template-matching method.

The position accuracy of our methodology (maximum error is less than 22 mm) performs similarly to previous studies [30,31] (maximum error was less than 25.4 mm) based on the two cameras. Note that the monocular camera was used, which is convenient for markers placed on a working device with a limited area.

Unlike previous studies [12,13,30,31], orientation accuracy was given in this study. However, it is an unsatisfactory result due to the relatively higher orientation error (the average error is about 4°), a possible reason could be the sensitivity of the rotation matrix. Estimating complete poses (i.e., position and orientation) could provide more comprehensive information for managing production in construction, such as productivity analysis, action recognition, and analysis of the interaction between works and entities. Therefore, it could be concluded that our method achieves a good performance and applicability for estimating the pose of an excavator.

Although the results show that the pose accuracy of the measurement system is sufficient in the construction industry, there are still some limitations that need to be addressed in future work. First, the system's effective measurement depth is not sufficient in long-distance work (e.g., land levelling, trenching). Thus, it is necessary to increase the working range without increasing the size of the marker.

A possible solution to overcome this limitation is to model the surrounding environment via LIDAR and register the laser point cloud with the image, which could increase the system's measurement depth. Second, a large number of markers are inconvenient to install on construction sites, which is time-consuming and labor-intensive. Therefore, excavator characteristics should be fully utilized for detection, such as articulated manipulator activity attributes.

For this limitation, semantic image segmentation technology could be implemented to track the excavator manipulator, which allows us to separate the manipulator from the background. Then the excavator key points will be extracted, and the pose is calculated using Equation (3). Lastly, similar to other vision-based detection technologies, marker occlusion or insufficient lighting (e.g., night, dense fog) will cause a system failure. More cameras could be installed to prevent occlusion, or other operating methods could be chosen for excavation operations.

As mentioned in the previous section, this paper intends to explore a form of monitoring technology based on monocular vision, which is flexible for layout under construction. The results of this study provide an alternative solution to the CV technology used in monitoring construction machinery.

6. Conclusions

In this work, the authors initially proposed an approach based on monocular vision to automatically estimate the excavator pose by detecting the marker installed on the excavator, which consisted of a marker with significant resistance to occlusion and a common monocular camera. Then, a measurement system was designed to estimate pose based on the pose estimation approach, and an error analysis was proposed to assess the accuracy of the system. Finally, some application guidelines were developed based on the results. The camera was attached to the side of the excavator (a convenient way to view multiple machines with the same camera at construction sites). This approach detected that the system depth was more than 11 m, the orientation error was less than 8.5°, and the position error was less than 22 mm, which could satisfy construction requirements. This prototype and experiments proved the effectiveness of the approach for practical application.

Future work needs to focus on the feasibility of the measurement system in the following aspects: (1) make full use of the excavator's manipulator structural characteristics and use semantic image segmentation technology to track manipulator movement in real-time, to improve the robustness of the system; (2) expand the working range of the measurement system and improve the versatility of the system; and (3) compare the performance of the measurement system under the different size of the marker. In addition, the fusion of high pixel camera multi-sensor information with inclination sensors could be considered as improving the level of automation of the excavator. Such a fully functional system would likely satisfy the requirements of construction applications.

Author Contributions: Conceptualization, J.Z., Y.H. and M.T.; methodology, J.Z., Y.H. and M.T.; software, J.Z.; validation, J.Z. and M.T.; formal analysis, J.Z.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z. and M.T.; supervision, Y.H. and M.T.; project administration, Y.H. and M.T.; funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities of Chang'an University, grant number 300102259503.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: The authors are gratefully thankful for the support by the Fundamental Research Funds for the Central Universities of Chang'an University (grant number 300102259503). The authors would like to thank the reviewers and the editor for their insightful comments, who helped improve the quality of this paper. The authors also thank graduate researchers Yanwei He, Xianglong Zhang, Yang Chen, Yang Ding, Lei Zhang, Xiaohua Xia, and Ni Zhao for their assistance in this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. BLS. Labor Productivity and Costs. Available online: https://www.bls.gov/lpc/construction.htm (accessed on 11 January 2021).
2. MOHURD. Announcement of the General Office of the Ministry of Housing and Urban-Rural Development on the Production Safety Accidents of Housing and Municipal Engineering in 2019. Available online: http://www.mohurd.gov.cn/wjfb/202006/t20200624_246031.html (accessed on 21 November 2020).
3. Bock, T. The future of construction automation: Technological disruption and the upcoming ubiquity of robotics. *Autom. Constr.* **2015**, *59*, 113–121. [CrossRef]
4. Seo, J.; Han, S.; Lee, S.; Kim, H. Computer vision techniques for construction safety and health monitoring. *Adv. Eng. Inform.* **2015**, *29*, 239–251. [CrossRef]
5. Liang, C.-J.; Kamat, V.R.; Menassa, C.M. Real-time construction site layout and equipment monitoring. In Proceedings of the Construction Research Congress 2018, New Orleans, LA, USA, 2–4 April 2018; pp. 64–74.
6. Cheng, T.; Mantripragada, U.; Teizer, J.; Vela, P.A. Automated trajectory and path planning analysis based on ultra wideband data. *J. Comput. Civ. Eng.* **2012**, *26*, 151–160. [CrossRef]
7. Pradhananga, N.; Teizer, J. Automatic spatio-temporal analysis of construction site equipment operations using GPS data. *Autom. Constr.* **2013**, *29*, 107–122. [CrossRef]
8. Mathur, N.; Aria, S.S.; Adams, T.; Ahn, C.R.; Lee, S. Automated cycle time measurement and analysis of excavator's loading operation using smart phone-embedded IMU sensors. In Proceedings of the Computing in Civil Engineering 2015, Austin, TX, USA, 21–23 June 2015; pp. 215–222.
9. Zou, J.; Kim, H. Using hue, saturation, and value color space for hydraulic excavator idle time analysis. *J. Comput. Civ. Eng.* **2007**, *21*, 238–246. [CrossRef]
10. Rezazadeh Azar, E.; McCabe, B. Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos. *Autom. Constr.* **2012**, *24*, 194–202. [CrossRef]
11. Rezazadeh Azar, E.; Dickinson, S.; McCabe, B. Server-customer interaction tracker: Computer vision–based system to estimate dirt-loading cycles. *J. Constr. Eng. Manag.* **2013**, *139*, 785–794. [CrossRef]
12. Yuan, C.; Cai, H. Key nodes modeling for object detection and location on construction site using color-depth cameras. In Proceedings of the Computing in Civil and Building Engineering, Orlando, FL, USA, 23–25 June 2014; pp. 729–736.
13. Yuan, C.; Li, S.; Cai, H. Vision-based excavator detection and tracking using hybrid kinematic shapes and key nodes. *J. Comput. Civ. Eng.* **2017**, *31*, 04016038. [CrossRef]
14. Soltani, M.M.; Zhu, Z.; Hammad, A. Skeleton estimation of excavator by detecting its parts. *Autom. Constr.* **2017**, *82*, 1–15. [CrossRef]
15. Soltani, M.M.; Zhu, Z.; Hammad, A. Automated annotation for visual recognition of construction resources using synthetic images. *Autom. Constr.* **2016**, *62*, 14–23. [CrossRef]
16. Xu, J.; Yoon, H.-S.; Lee, J.Y.; Kim, S. Estimation of excavator manipulator position using neural network-based vision system. *SAE Tech. Pap.* **2016**. [CrossRef]
17. Xu, J.; Yoon, H.-S. Vision-based estimation of excavator manipulator pose for automated grading control. *Autom. Constr.* **2019**, *98*, 122–131. [CrossRef]
18. Liang, C.-J.; Lundeen, K.M.; McGee, W.; Menassa, C.C.; Lee, S.; Kamat, V.R. A vision-based marker-less pose estimation system for articulated construction robots. *Autom. Constr.* **2019**, *104*, 80–94. [CrossRef]
19. Liang, C.-J.; Lundeen, K.M.; McGee, W.; Menassa, C.C.; Lee, S.; Kamat, V.R. Stacked Hourglass Networks for Markerless Pose Estimation of Articulated Construction Robots. In Proceedings of the 35th International Symposium on Automation and Robotics in Construction (ISARC), Berlin, Germany, 20–25 July 2018; pp. 859–865.
20. Liang, C.-J.; Lundeen, K.M.; McGee, W.; Menassa, C.C.; Lee, S.; Kamat, V.R. Fast dataset collection approach for articulated equipment pose estimation. In Proceedings of the Computing in Civil Engineering 2019, Atlanta, GA, USA, 17–19 June 2019; pp. 146–152.
21. Calderon, W.T.; Roberts, D.; Golparvar-Fard, M. Synthesizing Pose Sequences from 3D Assets for Vision-Based Activity Analysis. *J. Comput. Civ. Eng.* **2021**, *35*, 04020052. [CrossRef]
22. Fang, Q.; Li, H.; Luo, X.; Li, C.; An, W. A sematic and prior-knowledge-aided monocular localization method for construction-related entities. *Comput. Aided Civ. Infrastruct. Eng.* **2020**, *35*, 979–996. [CrossRef]
23. Luo, H.; Wang, M.; Wong, P.K.-Y.; Tang, J.; Cheng, J.C.P. Vision-Based Pose Forecasting of Construction Equipment for Monitoring Construction Site Safety. In Proceedings of the 18th International Conference on Computing in Civil and Building Engineering (ICCCBE 2020), São Paulo, Brazil, 18–20 August 2020; pp. 1127–1138.
24. Luo, H.; Wang, M.; Wong, P.K.-Y.; Cheng, J.C.P. Full body pose estimation of construction equipment using computer vision and deep learning techniques. *Autom. Constr.* **2020**, *110*, 103016. [CrossRef]
25. Luo, H.; Wang, M.; Wong, P.K.-Y.; Tang, J.; Cheng, J.C.P. Construction machine pose prediction considering historical motions and activity attributes using gated recurrent unit (GRU). *Autom. Constr.* **2021**, *121*, 103444. [CrossRef]
26. Rezazadeh Azar, E.; Feng, C.; Kamat, V.R. Feasibility of in-plane articulation monitoring of excavator arm using planar marker tracking. *J. Inf. Technol. Constr.* **2015**, *20*, 213–229.
27. Olson, E. AprilTag: A robust and flexible visual fiducial system. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 3400–3407.

28. Lundeen, K.M.; Dong, S.; Fredricks, N.; Akula, M.; Kamat, V.R. Electromechanical development of a low cost end effector pose estimation system for articulated excavators. In Proceedings of the 32nd International Symposium on Automation and Robotics in Construction and Mining (ISARC 2015), Oulu, Finland, 15–18 June 2015; pp. 1–8.
29. Lundeen, K.M.; Dong, S.; Fredricks, N.; Akula, M.; Seo, J.; Kamat, V.R. Optical marker-based end effector pose estimation for articulated excavators. *Autom. Constr.* **2016**, *65*, 51–64. [CrossRef]
30. Feng, C.; Dong, S.; Lundeen, K.M.; Xiao, Y.; Kamat, V.R. Vision-based articulated machine pose estimation for excavation monitoring and guidance. In Proceedings of the 32nd International Symposium on Automation and Robotics in Construction and Mining (ISARC 2015), Oulu, Finland, 15–18 June 2015; pp. 1–9.
31. Feng, C.; Kamat, V.R.; Cai, H. Camera marker networks for articulated machine pose estimation. *Autom. Constr.* **2018**, *96*, 148–160. [CrossRef]
32. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [CrossRef]
33. Fiala, M. ARTag, a fiducial marker system using digital techniques. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 590–596.
34. Atcheson, B.; Heide, F.; Heidrich, W. Caltag: High precision fiducial markers for camera calibration. In Proceedings of the 15th International Workshop on Vision, Modeling and Visualization (VMV 2010), Siegen, Germany, 15–17 November 2010; pp. 41–48.
35. Sagitov, A.; Shabalina, K.; Sabirova, L.; Li, H.; Magid, E. ARTag, AprilTag and CALTag fiducial marker systems: Comparison in a presence of partial marker occlusion and rotation. In Proceedings of the 14th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2017), Madrid, Spain, 26–28 July 2017; pp. 182–191.
36. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2004. [CrossRef]
37. Gao, X.S.; Hou, X.R.; Tang, J.; Cheng, H.F. Complete solution classification for the perspective-three-point problem. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 930–943. [CrossRef]
38. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EPnP: An accurate $O(n)$ solution to the pnp problem. *Int. J. Comput. Vis.* **2009**, *81*, 155. [CrossRef]
39. Penate-Sanchez, A.; Andrade-Cetto, J.; Moreno-Noguer, F. Exhaustive linearization for robust camera pose and focal length estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2387–2400. [CrossRef] [PubMed]
40. Triggs, B.; McLauchlan, P.; Hartley, R.; Fitzgibbon, A. Bundle adjustment—A modern synthesis. In Proceedings of the International Workshop on Vision Algorithms (IWVA 1999), Corfu, Greece, 21–22 September 1999; pp. 298–372.
41. Kümmerle, R.; Grisetti, G.; Strasdat, H.; Konolige, K.; Burgard, W. G2o: A general framework for graph optimization. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 3607–3613.