Automation in Construction 104 (2019) 80–94

Contents lists available at ScienceDirect

Automation in Construction

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

A vision-based marker-less pose estimation system for articulated construction robots

Ci-Jyun Liang[a], Kurt M. Lundeen[a], Wes McGee[b], Carol C. Menassa[a], SangHyun Lee[a], Vineet R. Kamat[a,*]

[a] Civil and Environmental Engineering, Univ. of Michigan, 2350 Hayward Street, 2340 G.G. Brown Building, Ann Arbor, MI 48109, USA
[b] Taubman College of Architecture and Urban Planning, Univ. of Michigan, 2000 Bonisteel Boulevard, Ann Arbor, MI 48109, USA

ARTICLE INFO

Keywords:
2D and 3D pose estimation
Stacked hourglass network
Human-robot collaboration
Construction safety

ABSTRACT

The prospect of human-robot collaborative work on construction sites introduces new workplace hazards that must be mitigated to ensure safety. Human workers working on tasks alongside construction robots must perceive the interaction to be safe in order to ensure team identification and trust. Detecting the robot pose in real-time is thus an essential requirement to inform the workers and to enable autonomous operation. Vision-based (marker-based, marker-less) and sensor-based are two of the primary methods for estimating robot pose. The marker-based and sensor-based methods require some additional preinstalled sensors or markers, whereas the marker-less method only requires an on-site camera system, which is common on today's construction sites. This research developed a marker-less pose estimation system for on-site articulated construction robots, which is based on a deep convolutional network human pose estimation algorithm: stacked hourglass network. Both 2D and 3D pose are estimated. The system is trained with image datasets collected from a robotic excavator and annotations of excavator pose, as well as conventional excavators working on construction sites. A KUKA robot arm with a bucket mounted on the end-effector was used to represent the robotic excavator in the experiments. The marker-less 3D method was evaluated, and the results were compared with the sensor-based results and the robot's ground truth pose. The results demonstrated that the marker-less 2D and 3D pose estimation methods are capable of performing proximity detection and object tracking on construction sites and can overcome the missing data issues encountered in the sensor-based method. However, the lower accuracy of the bucket pose estimation due to occlusion highlights the need for modifying the network and collecting additional datasets for training in future work.

1. Introduction

Due to the hazardous, unstructured, and dynamic working environment and labor-intensive nature, the construction industry has a higher rate of workplace fatalities and injuries compared to other industries [1,2]. According to reports from U.S. Bureau of Labor Statistics and CPWR, on average 53% of the fatal accidents that happen on construction sites are either struck by vehicle or equipment overturns and collisions between 2003 and 2010 [3], which costs approximately $13 billion per year in U.S. [4]. On a typical construction site, workers and heavy equipment have to work together closely, which increases the potential safety risks [5]. Blind spots around the equipment are the primary cause of such accidents [6]. When workers need to interact with the equipment on job sites, the equipment operator sometimes cannot locate all workers nearby and the workers also cannot monitor the equipment components clearly, especially for articulated equipment such as excavators that usually work around trenches or earth mounds that serve as potential occlusions leading to increased possibility of blind spots. In order to prevent these type of accidents, manual jobsite safety observations and inspections are required on construction sites [7]. However, safety personnel have to pay attention to entire jobsites continuously, which is time-consuming and incurs additional costs [8].

Underground utility strike incidents are another category of accidents related to the operation of articulated construction robot such as excavators [9,10]. According to the Common Ground Alliance (CGA) 2016 Damage Information Reporting Tool (DIRT) report, approximately 379,000 underground utility damage incidents were reported in 2016 in the U.S., which was an increase of 20% from 2015 and cost an

* Corresponding author.
E-mail addresses: cjliang@umich.edu (C.-J. Liang), klundeen@umich.edu (K.M. Lundeen), wesmcgee@umich.edu (W. McGee), menassa@umich.edu (C.C. Menassa), shdpm@umich.edu (S. Lee), vkamat@umich.edu (V.R. Kamat).

https://doi.org/10.1016/j.autcon.2019.04.004
Received 11 November 2018; Received in revised form 4 April 2019; Accepted 6 April 2019
Available online 18 April 2019

additional $1.5 billion [11]. One key reason for the high incident rate is the location uncertainty of the underground utilities [12]. Many of the existing buried utilities are abandoned or undocumented, and locating hidden utilities is the first step to address this issue [13]. The underground utility record could help workers and excavator operators avoid the potential utility locations. However, the operators sometimes cannot locate the bucket or utilities directly from the cabin. The indirect guidance from workers near the bucket does little to reduce the risks of utility strikes. Thus, utilizing sensors to estimate the excavator pose and providing real-time information to workers and operators has emerged as a feasible method and has been studied in developing on-site articulated construction robot pose estimation systems [12,14–16], and enhancing the on-site information with Augmented Reality [17–19]. Furthermore, the pose estimation system also provides the potential application of productivity analysis [20]. The existing productivity analysis methods only tracked the construction equipment or part of the equipment by sensors or computer vision method [21–23]. For example, the part of the excavator and the haul truck were identified and tracked during the dirt-loading cycle and utilized to estimate the productivity [21]. The motion analysis or action recognition methods are required to classify similar excavator activities such as digging and dumping to enhance productivity analysis [24]. This can be achieved by providing the detailed pose of the excavator for identifying the action [25].

The prospect of human-robot collaboration (HRC) on construction sites further heightens these proximity safety concerns [26]. Unlike HRC in typical manufacturing settings, the robot on the construction site has to maneuver around the unstructured environment to their next task location. The workplace of the robot changes dynamically based on their location, which is a challenge for HRC safety. According to standards ISO 10218-1, ISO 10218-2 and ISO/TS 15066, the safety of the HRC must be adhered to either by stopping the robot before human contact, or be controlled by regulating force and speed limits [27]. The recently developed dynamic safety system utilized human detection sensors and optical sensors to adjust the robot speed according to the detected human action and the protective distance [28]. However, the protective distance, or safety zone, has to be very large since the optical sensors only identify the difference between current frame and previous frame instead of tracking the robot's exact pose, which causes the poor utilization of space [27]. On the other hand, the robot's onboard sensors are often failed due to magnetic disturbance by artifacts (IMU) or signal blockage in an urban canyon (GPS) [29]. In addition, the articulated construction robot has arbitrary and expansive movement around the unstructured construction site and is difficult to make the construction site a structured environment [30]. This highlights the need for developing an effective on-site pose estimation system for articulated construction robot and human workers.

The experimental testbed of the construction articulated robot in this paper was an excavator since it is ubiquitous equipment on jobsite and has a large blind spot [6]. The pose of the excavator can be described as the angle between each component (boom, stick, and bucket) and the six degree-of-freedom (6 DOF) coordinates of each joint (cabin-boom, boom-stick, stick-bucket, bucket end-effector). Fig. 1 depicts a 2D pose estimation system. The pose of the construction equipment, such as an excavator, can be described as the angle between each component (boom, stick, and bucket) and the six degree-of-freedom (6 DOF) coordinates of each joint (cabin-boom, boom-stick, stick-bucket, bucket end-effector). In the 2D case, the pose is defined as the pixel-wise coordinate and angle (X, Y, θ), whereas in the 3D case, the pose is defined as the world coordinate and roll-pitch-yaw $(X, Y, Z, \phi, \theta, \psi)$. Fig. 2 illustrates the excavator side view with the kinematic chain and the corresponding parameters. The pose of each excavator joint can be calculated using the angle and lengths of each component by Forward Kinematics [31], or directly estimated by sensors or vision [12]. Therefore, determining the location of each joint and the angle between each link is the primary goal of articulated construction robot pose estimation.

Fig. 1. Illustration of the 2D on-site pose estimation system on a video frame for both articulated construction robot and human workers. Red lines are the estimated pose. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

1.1. Existing pose estimation methods

In current practice, two types of pose estimation methods are mainly used on construction equipment or human worker – these are non-visual sensor-based and vision-based pose estimation methods. For non-visual sensor-based pose estimation methods, sensors such as Inertial Measurement Unit (IMU), Global Positioning System (GPS), Wireless Local Area Network (WLAN), Radio Frequency Identification (RFID), and Ultra-Wide Band (UWB) are mainly deployed on construction equipment and construction sites. IMU sensors need to be mounted on excavator links to measure the angle [9,32–35], which suffers from drift issues over time and magnetic interference [36]. GPS is effective for outdoor use only and also suffers from the signal blockage in an urban canyon [29], which is not suitable for some indoor or urban construction sites. WLAN systems require significant amount of effort for calibration [37]. The accuracy of the WLAN estimation depends on the distribution of the access point [38]. RFID and UWB methods both require sufficient preinstalled tags and readers on equipment and infrastructure [39–42]. They generally suffer from missing data issues [16] and are inadequate for pose estimation [43]. Besides, most of these methods cannot provide orientation information directly, except for IMU sensors, and are thus not suitable for construction scenarios.

On the other hand, vision-based pose estimation methods are capable of analyzing position information as well as orientation information directly from input data, such as videos or point clouds [44]. These methods generally recognize construction equipment on site

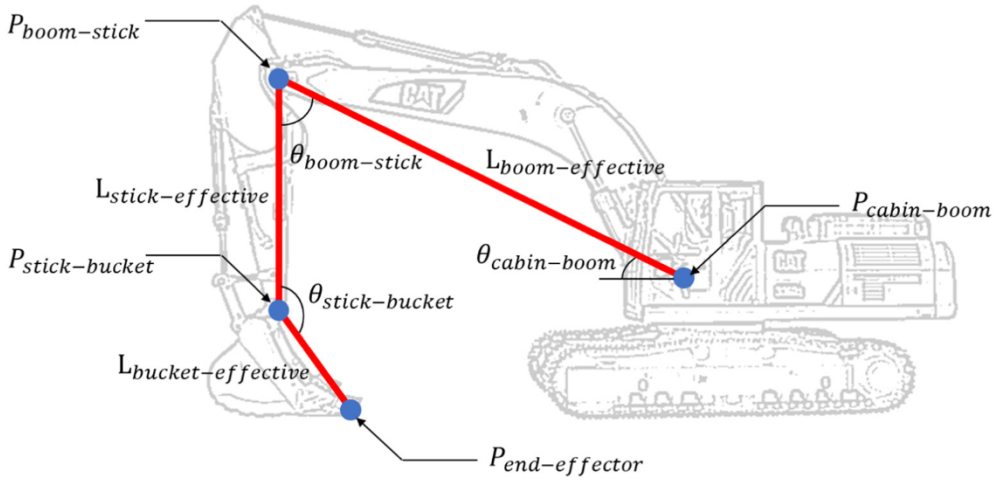

Fig. 2. Definition of the excavator pose. The pose of each joint P can be calculated by the component effective lengths L and the angle between each component θ.

[21,45–47], then estimate their six-degrees-of-freedom (6 DOF) pose [15,48,49], and can be categorized into two different groups: marker-based and marker-less pose estimation methods. The marker-based pose estimation method recognizes all the markers mounted on equipment and estimates the pose by their geometric relations or marker network [12,50,51], or projects infrared LEDs and analyzes the pattern to determine the pose [52,53], whereas the marker-less pose estimation method directly extracts image features and estimates the pose from them [15,48,49]. The marker-based method has been extensively applied in indoor localization and facility management [54–57]. Similar to the sensor-based pose estimation method, they also require pre-installed markers on equipment and environment.

In comparison to the marker-based method, the marker-less pose estimation method only requires an on-site camera system, which is common on typical construction sites today, or utilizes RGB-D cameras [58–61]. Feature descriptor based is the first type of marker-less pose estimation method, such as Histograms of Oriented Gradient (HOG) [21], 3D principal axes descriptor (PAD) [45], Iterative Closest Point (ICP) [62], or Viewpoint Feature Histogram (VFH) [48]. On the other hand, the recently emerging Convolutional Neural Networks (CNN) is another type of pose estimation method [63], which has improved performance (accuracy and speed) in comparison with all other vision-based methods, especially for human pose estimation. The majority of the human pose estimation methods are 2D-based methods [64,65], which estimate the human pose in 2D pixel-wise coordinates, as shown in Fig. 1. Existing human pose estimation can be categorized as detection-based and repression-based [66]. The detection-based methods utilized a heat-map to predict the joint location [67], whereas the regression-based methods utilized a nonlinear function to compute the joint coordinates directly [68]. The stacked hourglass network method proposed by Newell et al. [69] built the foundation of the state-of-the-art 2D human pose estimation method. Generative Adversarial Networks (GAN) [70], Pyramid Residual Module (PRM) [71], Conditional Random Field (CRF) [72] were applied to the stacked hourglass network to improve the performance. Besides, several existing 3D human pose estimation methods adopted the stacked hourglass network with coarse-to-fine volumetric architecture [73] or weakly-supervised approach [74]. The existing pose estimation method were mainly focused on the 2D pose due to the lack of 3D ground truth posture data [75]. For human pose data collection, the motion capture system is primarily used to obtain the ground truth data of human skeleton in an indoor environment [76], which is difficult to employ for construction equipment in an outdoor environment.

1.2. Applications of pose estimation methods

The existing pose estimation methods used in construction have different target applications. The accuracy and the specific shortcomings of any pose estimation method affect the method selected for each specific construction application. Table 1 lists the accuracy and the limitations of the existing pose estimation methods. For the 3D marker-less vision-based pose estimation method, the accuracy can be achieved at 1 m. However, the largest distance of the target equipment from the camera is 50 m; otherwise, the accuracy drops dramatically [15]. For the 3D marker-based vision-based pose estimation method, the accuracy can be achieved at 2 cm when the distance between camera and bucket teeth is under 6.1 m [12]. The camera occlusion is the main drawback of the marker-based method since the markers have to be visible in the camera view at all times in order to estimate the pose [12]. For the sensor-based pose estimation method, the accuracy can be achieved at 5 cm when testing on a real excavator arm with IMU sensors [33] but could be improved depending on the type of the sensor used. In addition, data missing or signal block is the major issue of the sensor-based method [16,33]. Finally, for the 2D vision-based pose estimation method, the angular accuracy can be achieved at 10° between the excavator components, which results in 122 cm vertical displacement when the reaching length of the excavator boom is 7 m [49]. However, this type of method can only provide 2D pixel-wise location or angle in each image and requires extra post-processing to acquire the depth data or 3D pose [49].

Pose estimation methods have been applied on construction sites to address safety and quality related issues. Table 2 compares the different pose estimation related construction applications comparing their acceptable location uncertainty and the methods currently used. The first application is preventing accidental utility strikes during excavation, which has a 2.5 cm acceptable location uncertainty [12]. The sensor-

Table 1
Comparison of the existing pose estimation methods by accuracy and limitations.

	3D marker-less vision-based [15]	3D marker-based vision-based [12]	Sensor-based [33]	2D vision-based [49]
Accuracy	1 m	2 cm	5 cm	10°
Disadvantage	Distance < 50 m	Camera occlusion	Data missing	No depth data and 3D pose

Table 2
Comparison of equipment pose estimation applications in the construction industry by location uncertainty and corresponding methods.

	Preventing utility strikes	Grade control	Object detection and tracking	Proximity detection	Autonomous excavation
Location uncertainty	2.5 cm [12]	2.5 cm [12]	< 1 m [79]	< 0.7 m [81]	4 cm [83]
Methods	Sensor [16,33]	Sensor [77,78]	Sensor [79]	Sensor [39,81]	Sensor [83,84]
	3D vision [12]		2D vision [14,80]	2D vision [82]	3D vision [15]

based method [16,33] and the 3D marker-based vision-based method [12] are two methods used for such applications. The second application is grade control, which also has a 2.5 cm acceptable location uncertainty [12]. Several sensor-based grade control commercial products have claimed that their accuracy can approach 1 mm [77,78]. The above two applications can tolerate relatively low uncertainty in pose estimates due to their precise control features.

The third application is object detection and tracking. The object detection and tracking methods have demonstrated a location uncertainty of < 1 m [79], and sensor-based methods and 2D vision methods are mainly utilized in this application [14,80]. The fourth application is proximity detection in which the location uncertainty is shown to be under 0.7 m [81]. Similar to object detection and tracking, the sensor-based method [39,81] and the 2D vision method [82] are used in proximity detection applications. Instead of the high accuracy, the data consistency is more important for these two types of applications. Finally, the fifth application is autonomous excavation, and the acceptable location uncertainty is 4 cm [83]. The sensor-based method [83,84] and the 3D vision-based method [15] are applied.

2. Research goal and contribution

In this study, a vision-based marker-less pose estimation system for articulated construction robots is proposed, which can distinguish robot joints and estimate their poses in images or video frames. The excavator is used as the experiment testbed. This system is built on a state-of-the-art human pose estimation deep neural network called the stacked hourglass network [69,85] and trained on an excavator image dataset collected from a factory environment with a robotic manipulator. The network is adapted and modified for the excavator skeleton. Both 2D and 3D versions of the system are built and evaluated in order to characterize the location uncertainty requirement illustrated in Table 2. The performance of the proposed system is validated based on the dataset annotation and the ground truth data, and compared with the sensor-based pose estimation method (IMU sensors). In addition, a fast dataset collection approach for articulated construction robot pose estimation is also developed and described.

The remainder of this paper is organized as follows. First, a deep neural network vision-based pose estimation method for articulated construction robot is introduced. Both 2D and 3D version baselines are established and evaluated. Second, the performance of the proposed pose estimation method is investigated via an experiment and compared with the IMU-based pose estimation method. Lastly, an articulated construction robot pose estimation dataset is collected and evaluated.

3. Vision-based marker-less 2D pose estimation

The proposed vision-based marker-less 2D pose estimation system is developed based on a state-of-the-art human pose estimation algorithm, namely the stacked hourglass network by Newell et al. [69,85]. This network scales the training images into different resolutions and captures features, and then combines the information to predict the pose. Compared to the complicated human pose, the construction equipment pose is relatively simpler and thus requires less information across different image resolutions. The detailed network architecture is further discussed in the next section.

3.1. 2D pose estimation system network architecture

Unlike the complicated human skeleton, excavator pose only requires identifying three components, which are the bucket, stick, and boom, and their corresponding joints. Therefore, the complexity of the network needed is lower than the original network. Two convolutional layers followed by a max pooling layer are first applied to the training images, which shrinks the images down to the size of 64 pixels. Then three subsequent convolutional layers upscale the images to the size of 256 pixels before the hourglass module. Finally, four hourglass modules, output prediction modules, and residual link modules are used in the network. According to Newell et al. [69], eight hourglass modules are used for human pose estimation. The reason for using four hourglass modules for the excavator pose estimation is that the excavator pose is relatively simpler than the human and thus requires less information across different image resolutions. All the convolutional layers are followed by ReLu activation function, with stride 1 except the first convolutional layer (Conv1 layer) with stride 2, and with batch normalization except the convolutional layers in the output prediction module. Fig. 3 shows the detailed network architecture.

The hourglass modules are the main components that collect features across different resolution of the images. Fig. 4 shows the network architecture of the hourglass module. The input passes into two parallel routes. In the first route, only one convolutional layer is applied to

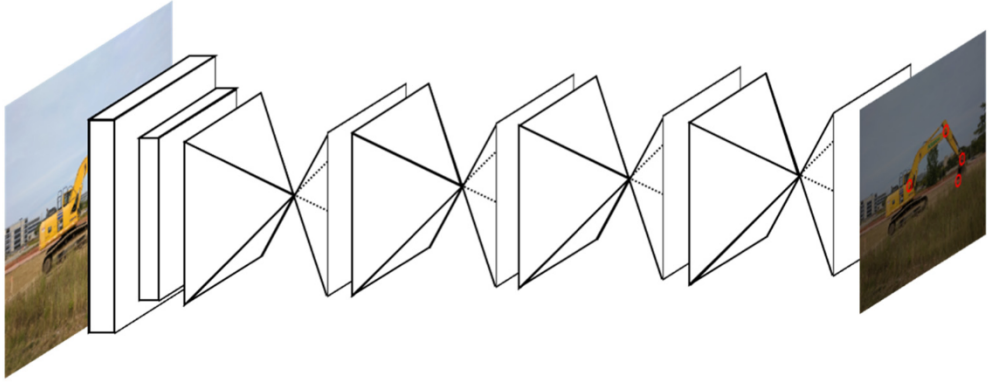

Fig. 3. Vision-based marker-less 2D pose estimation network architecture [69].

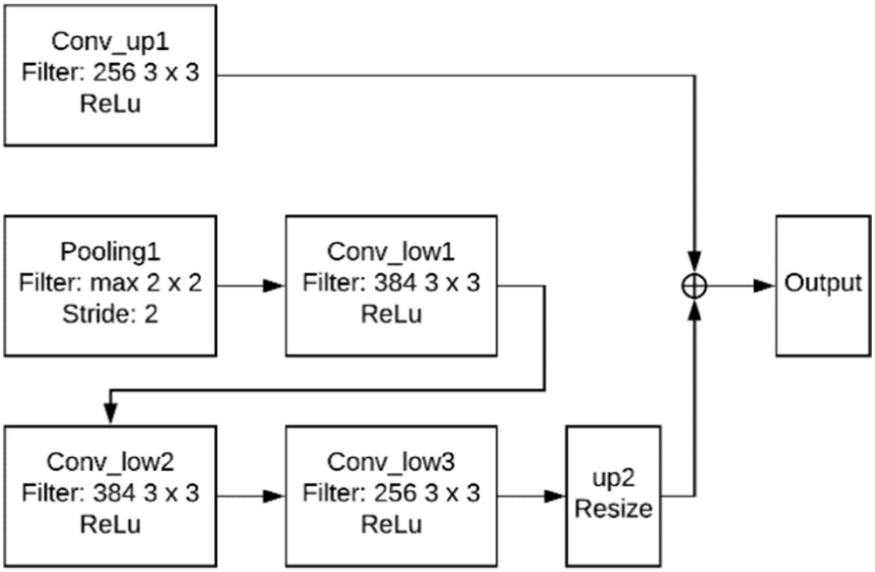

Fig. 4. Hourglass module network architecture.

upscale the input to the size of 256 pixels. In the second route, one max pooling layer followed by three convolutional layers are applied to downscale the input to the size of 384 pixels, then resized to the size of 256 pixels, as the first route result. Finally, two route results are added together through elementwise summation to generate the output. This can preserve the global features and capture the local features as well.

The output prediction module and residual link module are applied after the hourglass module. Two convolutional layers are used in the output prediction module to generate the heat-map of the possibility distribution of the location of each joint. Fig. 5 shows the concept of the prediction heat-map. Each circle in the image represents the highest probability of the corresponding joint location. The final layer is a one-by-one convolutional layer, which aims to calculate the possibility across the depth of the output of the previous layer. On the other hand, the residual link module combines the output of the prior hourglass and the output prediction module to generate the input for the next hourglass. The repeated hourglass and residual link modules can preserve the spatial location and relation of each feature and apply to the final prediction step.

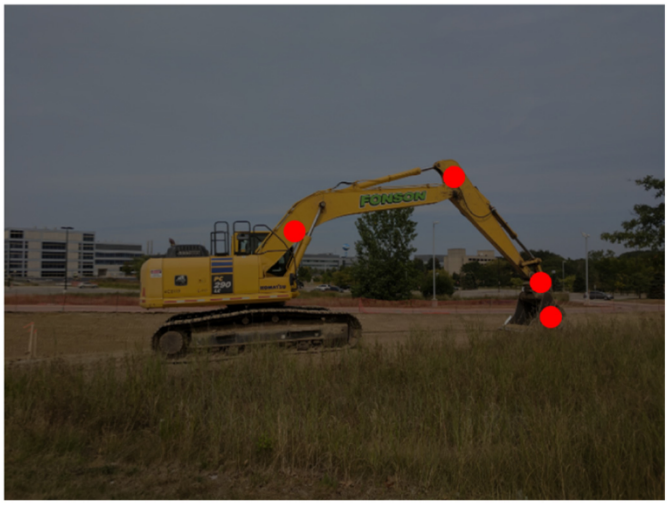

Fig. 5. The concept of the prediction heat-map generated by the output prediction module. Each circle represents the highest probability of corresponding joint location.

3.2. Training details and implementation

The L_2-norm loss function is used to train the network, as shown in Eq. (1):

$$L_2(\widehat{X}_P, X_L) = \sum (\widehat{X}_P - G(X_L))^2 \quad (1)$$

where $\widehat{X}_P$ represents the predicted pose and X_L represents the labeled ground truth training data, and $G(\cdot)$ represents the Gaussian kernel function with 1-pixel standard deviation. The loss function directly calculates the error between the training ground truth heat-map and the predicting heat-map and minimizes it.

The network system is implemented by modifying the original network using PyTorch and the loss function described above. The RMSprop method with learning rate 2e-4 is used for optimization. Batch normalization is used for the training process [86]. The network is trained with NVIDIA GeForce GTX 1060 graphic card on an excavator image dataset, which is collected from a factory setup laboratory environment with a simulated robotic excavator. The excavator dataset contains 3000 training images and 500 testing images aligned with their 2D pose annotation. The detailed laboratory environment setup and data annotation are discussed in Sections 5.1 and 5.2.

4. Vision-based marker-less 3D pose estimation

The proposed vision-based marker-less 3D pose estimation system is adapted and modified from a 3D human pose estimation baseline network [75]. This network uses the 2D pose estimation result, such as the stacked hourglass network, to predict and reconstruct the 3D pose. This can expedite the estimation process in order to accomplish the real-time pose estimation. The detailed network architecture is illustrated in the following section.

4.1. 3D pose estimation system network architecture

The objective of the baseline network is to predict and reconstruct the 3D pose of the articulated equipment based on the input 2D pose data. The 2D pose data from the previous vision-based marker-less 2D pose estimation result is passed to two subsequent linear layers, which are followed by the ReLu activation function and 0.5 dropout. The batch normalization is also applied to the linear layer output, which can increase the performance of the network. Next, the residual link module combines the output of the linear layer and the input 2D pose data to

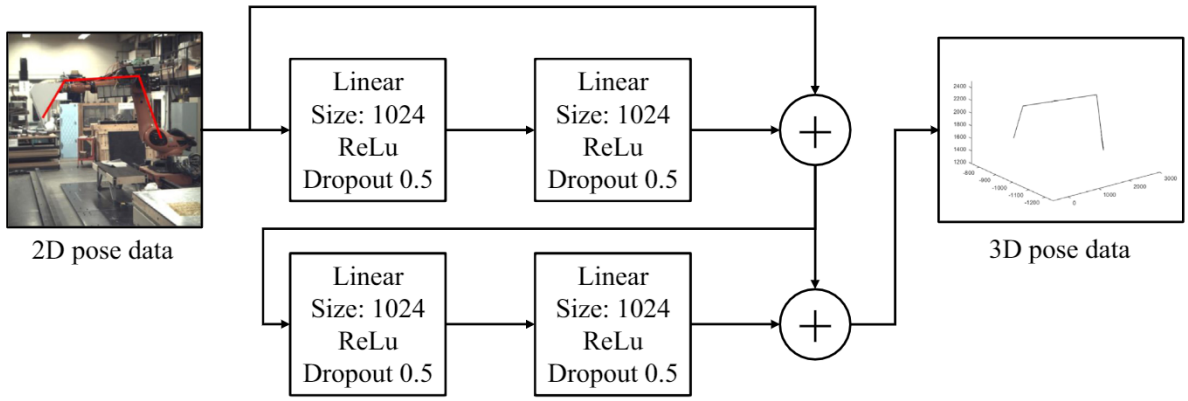

Fig. 6. Vision-based marker-less 3D pose estimation network architecture. Two linear layers and residual link are repeated twice to reconstruct the 3D pose based on the input 2D pose [75].

generate the predicted 3D pose, similar to the 2D pose estimation network. The entire process is repeated twice to generate a higher accuracy of the prediction and prevent overfitting. Based on the experiment results from [75], the best performance of the network can be achieved by repeating the process twice and it will saturate after repeating the process four times due to overfitting the network. Fig. 6 shows the 3D pose estimation network architecture.

4.2. Training details and implementation

The L_2-norm loss function is also used to train the network, as shown in Eq. (2):

$$L_2(X_{2D}, X_{3D}) = \sum (f(X_{2d}) - X_{3D})^2 \tag{2}$$

where X_{2d} represents the input 2D pose data, and X_{3D} represents the labeled ground truth 3D training data, and $f(\cdot)$ represents the function that maps the 2D input data to the 3D prediction. The loss function minimizes the prediction error between 3D prediction and 3D ground truth data. The L_2-norm loss function is derived from the loss function of the 3D human pose estimation baseline network [75].

The network is implemented using TensorFlow, and the loss function described in Eq. (2). The Adam method with starting learning rate 2e-3 and exponential decay is used for optimization, instead of starting learning rate 1e-3 [75]. Batch normalization is also used for the training process. The network is trained with NVIDIA GeForce GTX 1060 graphic card on the same image dataset collected from the laboratory with a robotic excavator. The 3D ground truth data is measured directly from the robot's embedded joint sensors. The complete laboratory setup is discussed in Section 5.1.

5. Dataset collection

The image dataset is collected with an articulated robotic manipulator outfitted with a simulated excavator bucket. The dataset is separated into training and testing groups. The proposed networks are trained by the training group and then evaluated by the testing group.

5.1. Dataset collection setup

For the dataset collection setup, a KUKA seven DOF robot arm (KUKA KR120) [87] was used to simulate the excavator, and the images of the robot arm with different poses were captured. Fig. 7 illustrates the simulated excavator in the laboratory. The upper arm represents the excavator stick and the lower arm represents the excavator boom. A bucket is mounted on the robot arm end-effector for a more realistic simulation. In order to control the robot as an excavator, the profile of the mounted bucket must remain perpendicular to the ground level.

Fig. 7. The simulated robotic excavator - robot arm mounted with an excavator bucket.

Thus, only four of the robot joints were moved during the dataset collection process, and the others were fixed at all times. The robot arm was controlled to follow trajectories to perform several excavator-like tasks such as digging, swinging, or unloading. The ground truth of the excavator pose data was acquired from the robot arm's embedded encoders, including 6 DOF pose of the robot's end-effector (X, Y, Z, A, B, C) and angles of all joints $(A_1, A_2, A_3, A_4, A_5, A_6)$.

In order to collect the images of the simulated excavator, a Point Grey camera [88] was used in the process. The camera was mounted on a second KUKA robot arm in the laboratory, as shown in Fig. 8. This could not only provide several different locations and orientations of the camera to increase the variety of the dataset, but also helped obtain the 6 DOF pose of the camera itself, which is the end-effector of the camera robot, for further processing. The mounted camera on the second robot arm was triggered by the same controller (Programmable Logic Controller, PLC) utilized to control the first robot arm. Thus, the captured image and the recorded ground truth pose data were synchronized with each other. In the data collection process, a total of 2500 images were collected; 2000 of them were used as training images and 500 of them were used as testing images. The data augmentation method was applied to increase the verity of the dataset to 3000 training images [89]. The human pose benchmark dataset FLIC [90] is composed of 3987 training images. In addition, the human pose is much more complicated than the excavator pose and constraint-free. The excavator has 1 DOF joints which are finite in number, and that reduces very dramatically the number of images needed for training. Fig. 9

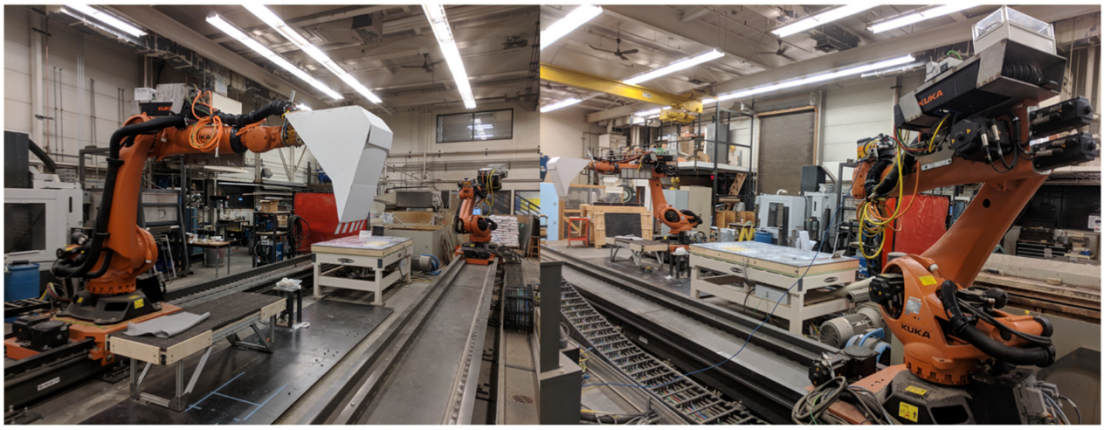

Fig. 8. The camera mounted on the second robot arm to capture the images.

shows a set of the collected images from the dataset. The size of each image is 2048 × 2048 pixels.

In addition, to increase the variety of the dataset and vary the background of the dataset images, several images from outdoor construction sites with working excavators were also collected, as shown in Fig. 10. The images contain the variety of excavator operations on different construction sites with single or multiple machines. These images were only used for evaluating the 2D pose estimation network since the 3D ground truth data could not be obtained from these images. A total of 500 images were collected; 400 of them were used as training images and 100 of them were used as testing images. The size of the images was different, and thus needed rescaling and cropping to 1024 × 1024 pixels before inputting into the network.

Fig. 9. A set of the captured images for the excavator dataset with different camera location and orientation, and excavator pose.

Fig. 10. A set of working excavator images from the dataset.

5.2. Data annotation

Data annotation is required in order to indicate the location of the excavator's joints in the images as the ground truth. The structure of the excavator data annotation follows the similar structure to the human pose dataset annotation, MPII for 2D pose [63] and Human3.6 M for 3D pose [76]. In the 2D pose annotation, excavator joint locations were annotated in the pixel-wise coordinate. The visibility of each joint was also marked in the annotation data. The scale of the image was measured with respect to a height of 200 pixels. On the other hand, in the 3D pose annotation, the locations of the excavator's joints were labeled as (X, Y) in pixel-wise coordinates and Z was considered as the depth value from the camera to each joint, which was calculated from the robot arm end-effector and joints' ground truth data. The bounding box was also labeled to show the area of the excavator in the image. The annotations were performed via MATLAB and saved as two separated annotation files, one for the 2D pose and the other for the 3D pose. Fig. 11 shows an example of an annotated image.

The 3D ground truth data was acquired by the robot arm's built-in encoders and the Programmable Logic Controller (PLC). Fig. 12 illustrates the framework of the pose data and image acquisition. The PLC sent the control command to both robot arms (North and South). The South robot would perform the predefined trajectory, such as digging or unloading, whereas the North robot would stay as it is to capture the images. Several trigger points were set to trigger the camera on the North robot to capture the image and acquire the pose of both robots, and then transfer them to a computer. After the South robot finished the entire trajectory, the North robot would move to the different pose and re-run the process. This could increase the variety of the dataset by having different orientations in the images. The 3D pose of the end-effector was directly read from the robot arm, and the 3D pose of the rest of the robot joints was obtained using inverse kinematics.

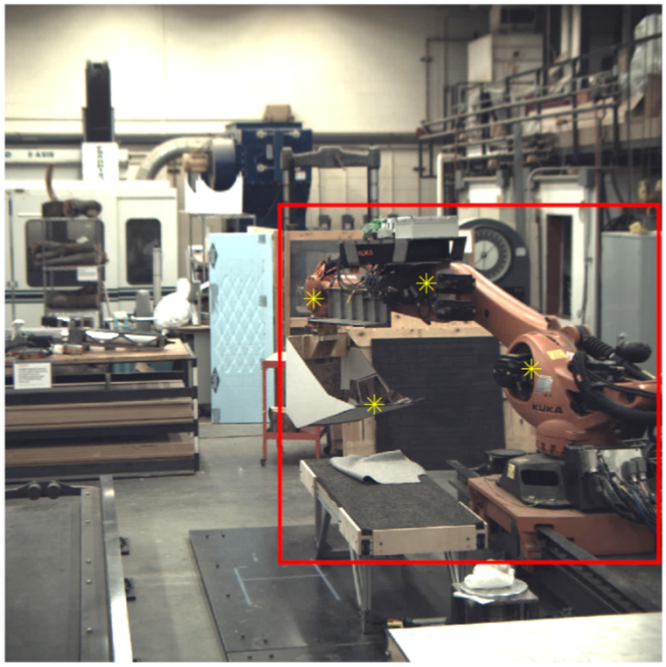

Fig. 11. An example of the annotated image. Stars represent the joint locations and the rectangle represents the bounding box.

5.3. Sensor-based pose estimation

For evaluating the vision-based pose estimation method, the sensor-based pose estimation method was used to compare the performance. Four IMU sensors were deployed to measure the angular change of the robot joints, as shown in Fig. 13. These sensors were placed on the axis of each joint so that they can measure the correct angle when the robot

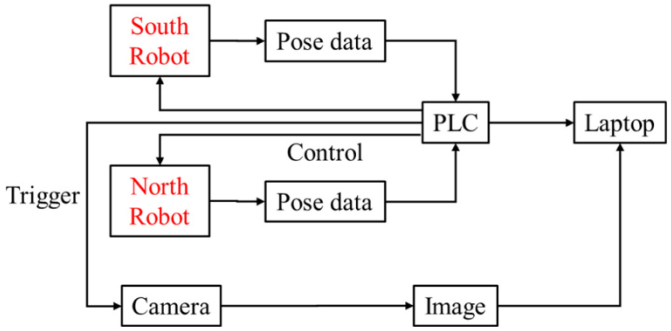

Fig. 12. The framework of the pose data and image acquisition.

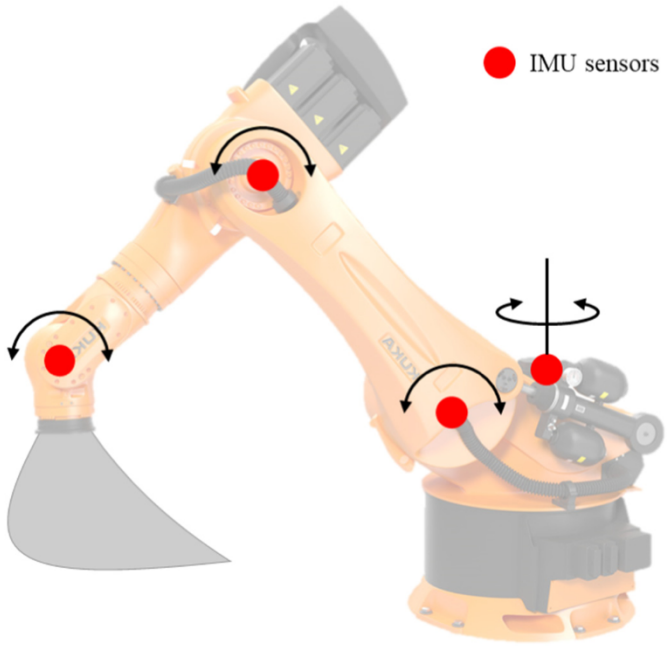

Fig. 13. IMU sensors deployment. Only two types of orientations are considered in the excavator pose.

changed its pose. The results were compared with the ground truth joint angle of the robot arm. The 3D pose of each joint could be calculated by forward kinematics. Since the exact location of a joint required location sensors such as GPS, which was not available in the experiment, the first joint (A1) was aligned with the ground truth A1 joint location, and then the other joints were calculated relative to the first joint. The Xsens MTw Awinda wireless motion tracker system [91] was used for the sensor-based method. The system contained four motion trackers with IMU embedded and a wireless receiver to transmit the data. The sensor data was also synchronized with the vision-based pose data and the ground truth data, so that it could be compared with each other. The results of the sensor-based method and the vision-based method are presented in the next section.

6. Experimental results

The results of the pose estimation experiments are explained in the following sub-sections. The 3D vision-based method, sensor-based method, and ground truth are compared with each other.

6.1. Results of the marker-less 2D pose estimation

The proposed 2D network was evaluated by comparing the prediction results of the testing images and the ground truth. Fig. 14 demonstrates the results of the excavator pose estimation. The lines represent the bucket, stick, and boom prediction. These images are estimated in the testing dataset. The Euclidean distance between the estimated joint location and the ground truth joint location are used to evaluate the performance, and the error percentage of the predicted component length and the ground truth, which can be seen in Tables 3 and 4. The average Euclidean distance between the laboratory testing dataset and ground truth is 40.64 pixels (image size is 2048 × 2048) and between the real site testing dataset and the ground truth is 71.84 pixels (image size is 1024 × 1024). In the laboratory dataset, the pixel size is measured by averaging the length of the robot arm across the entire dataset, which resulted in 1 pixel approximated to 1 mm. Therefore, the average Euclidean distance in the laboratory dataset can be converted to 40.64 mm. The distance between the camera and the robotic excavator is 10 m. The distance between the camera and the robotic excavator is 10 m. In the real site dataset, the distance between the camera and the excavator is unknown for each image since they were collected randomly online or on jobsite. Thus, it is difficult to convert the result from pixel to mm. The result is roughly converted to mm by measuring the length of the excavator stick in the testing image and calculating the ratio with actual excavator stick length. The size of the excavator must be similar throughout the entire testing dataset. The stick size in the testing image is 40 pixels and the actual stick size is 2500 mm, which resulted in 1 pixel approximated to 12 mm. Therefore, the average Euclidean distance in the rea site dataset can be converted to 862.08 mm.

The result showed that the bucket location has the highest error because the bucket is blocked (occluded) or out of range in some of the images. The network still tries to find the bucket location in these cases, which increases the error distance. The error in the real site dataset is higher than the laboratory dataset. This is because the real site dataset has a greater variety of excavators and backgrounds. Only some of these variations were included in the testing dataset, so this caused a decrease in accuracy. The number of images in the real site dataset is also insufficient for training purposes.

For the error percentage of the predicted component length and the ground truth, only the laboratory dataset was evaluated because the length of each robot arm skeleton is known, but some of the component sizes in the real site dataset are unknown because of occlusion. The results are shown in Table 4. The error percentage of the boom and stick is approximately 40% and 31%, and the bucket is 59%. The reason for the high error percentage in the bucket case is the occlusion issue. When the bucket is blocked or out of range in the image, the predicted bucket location will be far away from its actual location. In addition, the ground truth length of the bucket is short, which increases the differences between the ground truth and the false predicted result. Fig. 15 shows the result of a false prediction of the bucket caused by occlusion. The excavator is partially blocked by another equipment, and the network mispredicts the bucket pose.

6.2. Results of the marker-less 3D and the sensor-based pose estimation

The proposed 3D pose estimation method was first evaluated by comparing the prediction results and the ground truth of the laboratory dataset. Fig. 16 shows the result of the 3D pose estimation. The left image was the result of the 2D pose estimation, which was the input to the 3D network. The right image was the 3D predicted result. The dashed line is the vision-based result, the dotted line is the sensor-based result, and the solid line is the ground truth. The Euclidean distance between the estimated joint location and the ground truth joint location are used to evaluate the performance, as shown in Table 5. Since the boom location was aligned together, it was not considered in the comparison.

The average Euclidean distance between the 3D vision-based method and ground truth is 144.65 mm (distance between the camera and the robotic excavator is 10 m), and between the sensor-based

Fig. 14. The result of the excavator 2D pose estimation. On the top is the simulated excavator and on the bottom are real excavators. Lines represents the estimated pose.

Table 3
Results of the average Euclidean distance (mm) between the predicted and the ground truth joint location.

(mm)	Laboratory dataset	Real site dataset
Boom	31.58	777.12
Boom stick	39.47	701.40
Stick bucket	35.65	753.36
Bucket	55.84	1216.44

Table 4
Results of the error percentage of the predicted component length and the ground truth in the laboratory dataset.

(%)	Error percentage of the component length
Boom	39.1
Stick	30.7
Bucket	58.8

method and the ground truth is 93.66 mm. The result showed that the error of the 3D vision-based is higher than the sensor-based method. One of the reasons was that the 3D vision-based method predicted the pose based on the 2D pose estimation result, wherein the error would accumulate from 2D prediction and decrease the accuracy in the 3D prediction. The other reason was that the camera coordinates pre-processing mentioned in [75] was not applied to the ground truth data because the camera matrix was not determined in the laboratory

Fig. 15. False prediction result of the bucket due to occlusion.

dataset. In addition, the occlusion issue also affected the prediction result similar to the 2D results. The error caused by the occlusion also accumulated from the 2D pose estimation results, especially for the bucket.

Second, the bucket pose estimation accuracy was evaluated by comparing the estimated bucket location with the sensor-based result and the ground truth. In the laboratory dataset, a sequence of the

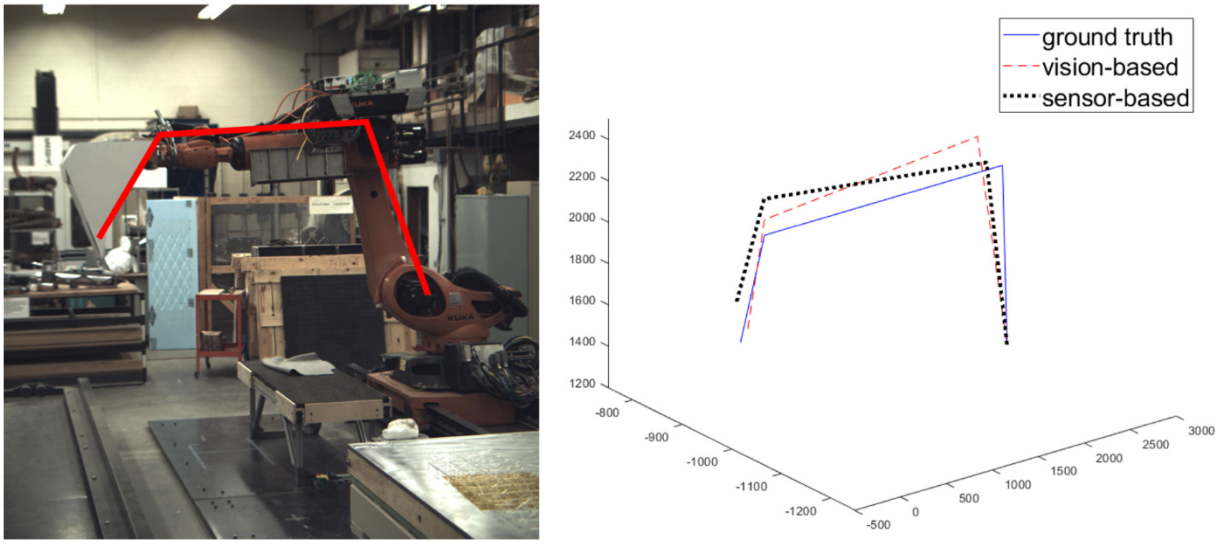

Fig. 16. Results of the excavator 3D pose estimation. The left image is the result from 2D pose estimation and the right image is the 3D result. The dashed line represents the vision-based result, the dotted line represents the sensor-based result, and the solid line represents the ground truth.

Table 5
Results of the average Euclidean distance (mm) between the predicted and the ground truth joint location.

(mm)	3D vision-based	Sensor-based
Boom	–	–
Boom-Stick	148.16	84.35
Stick-Bucket	134.22	97.21
Bucket	151.58	99.42

excavator trajectory was repeated ten times and was captured with different camera orientations, as demonstrated in Fig. 17. A total of 16 images were captured in the trajectory yielding a total of 160 images that were used in the evaluation. The average pose of each of the 16 data points was calculated and compared between pose estimation methods. Fig. 18 shows the results of the pose estimation.

The star-line is the 3D vision-based result, the circle-line is the sensor-based result, and the cross-line is the ground truth. The error of the 3D vision-based method is larger than the sensor-based method at the beginning of the trajectory. The sensor-based pose is closer to the ground truth pose than the vision-based pose before data 5 in X and Y location. After data 6, the sensor-based pose has a higher error than the vision-based pose. The error of the X and Y location in the sensor-based result increased over time. The difference in the Z location in sensor-based and vision-based pose does not change significantly. This is because the drift occurred in the heading direction (Yaw). The sensor system had a stabilizing mechanism to calibrate the sensors. The earth's magnetic field was used to stabilize the heading, but is susceptible to disturbance by artifacts such as nearby metal objects. In addition, the cumulative error of the bucket 3D vision-based pose estimation is illustrated in Fig. 19. The straight line is the error in X-axis, the cross-line is the error in Y-axis, and the circle-line is the error in *Z*-axis. The cumulative error along the X- and Y-axes is higher than the cumulative error along the Z-axis since the movement of the excavator bucket in the data points is larger in the X and Y direction. In addition, the cumulative error along the X-axis is much higher than along the Y- and Z-axes. This is because the X direction has a higher projection in the camera viewing direction and the movement in such direction is difficult to identify by a single camera. Moreover, the Z direction is tangent to the viewing direction of the camera (pointing up), which has a larger displacement in the image and results in better performance.

Third, the proposed 3D vision-based pose estimation method was evaluated by comparing the accuracy with the existing vision-based pose estimation method. In Table 1, the accuracy of the 3D marker-less vision-based method is 1000 mm (camera distance 50 m) [15], and the accuracy of the 3D marker-based vision-based method is 20 mm (camera distance 6.1 m) [12], whereas the accuracy of the proposed method is 144.65 mm (camera distance is 10 m). The results showed that the proposed method could achieve higher accuracy than the existing 3D marker-less method but the camera distance is shorter and

Fig. 17. A sequence of the excavator trajectory repeated ten times with different camera orientations.

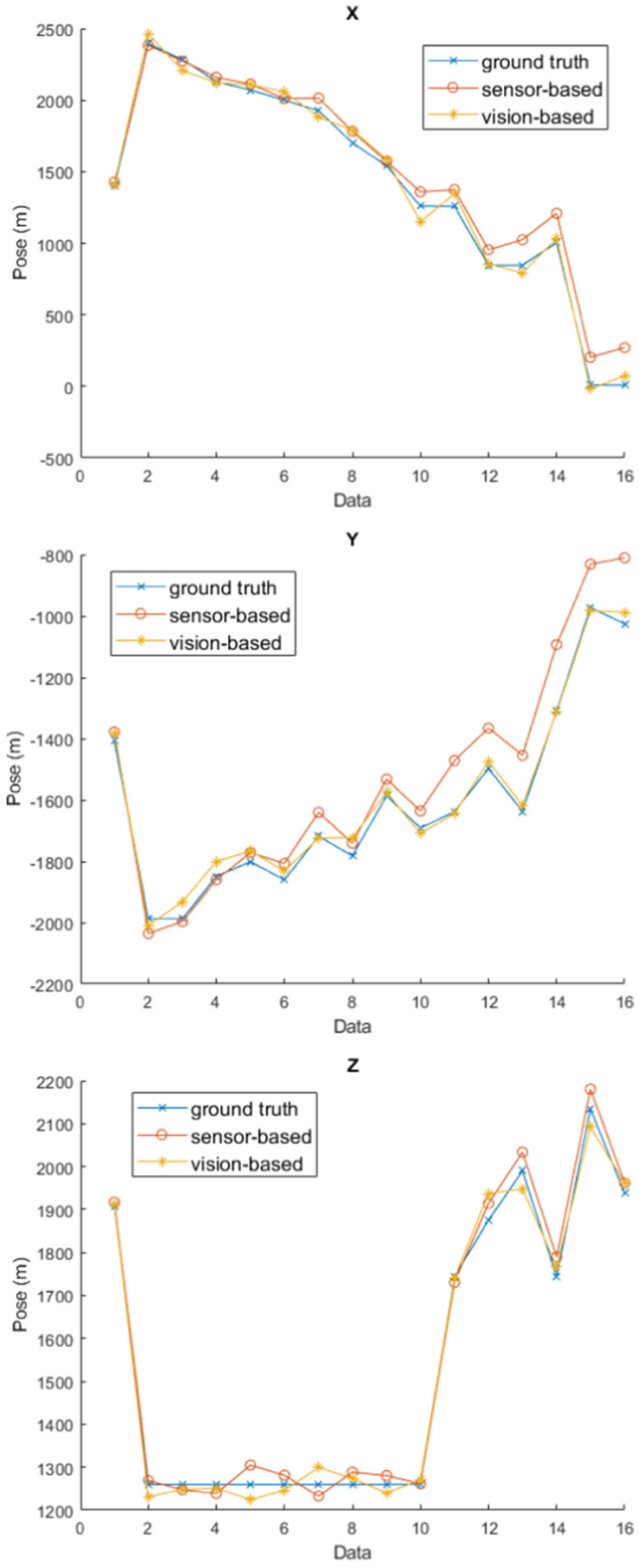

Fig. 18. Results of the bucket 3D pose estimation. The star-line is the 3D vision result, the circle-line is the sensor result, and the cross-line is the ground truth.

had lower accuracy than the 3D marker-based method. Even though the proposed method still has the occlusion issue, which is the standard issue of existing vision-based methods, it can be easily addressed by providing sufficient occlusion training data or dataset augmentation method.

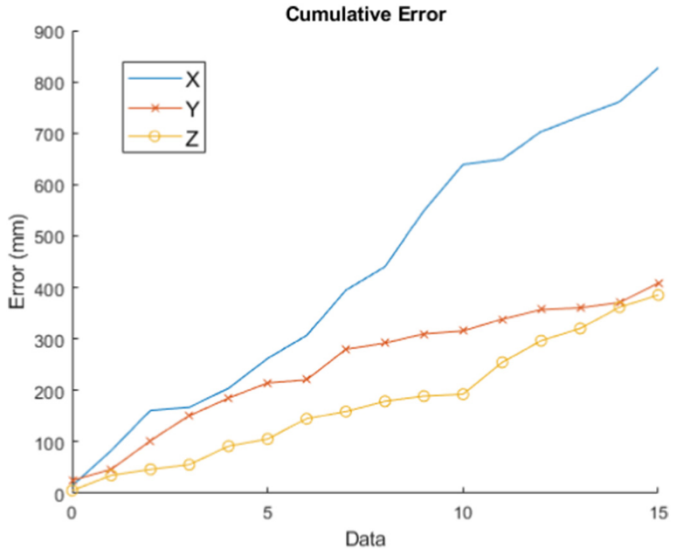

Fig. 19. Cumulative error of the bucket 3D vision-based pose estimation. The straight line is the error in X-axis, the cross-line is the error in Y-axis, and the circle-line is the error in *Z*-axis.

Finally, the processing time for completing 500 testing images is 416 s and averaging 0.832 s per image (1.2 Hz) with NVIDIA GeForce GTX 1060 graphic card. In general pose estimation practice, the real-time performance is defined as greater than or equal to 1 Hz [92], where the proposed method is above the threshold and thus can achieve the real-time performance, especially for the slow-moving articulated construction robot such as the excavator. The Pose Interpreter Networks method is 20 Hz for object pose with RGB-D camera and NVIDIA GeForce Titan X graphic card [93], the VNect method is 30 Hz for 3D human pose with RGB-D camera and NVIDIA GeForce Titan X graphic card [94], and the Part Affinity Fields method is 8 Hz for 2D multiple human with a single camera and NVIDIA GeForce 1080 graphic card [95]. The processing time of the proposed method can be improved by using advanced hardware.

7. Discussion

Based on the evaluation results, occlusion is the primary issue of the proposed vision-based method, which can potentially be addressed by increasing the number and variety of the training dataset. Dataset augmentation and expansion techniques can also help address this issue. Another problem is the multiple-machine situation. The proposed network can only identify one machine's pose. If there are two or more articulated machines in the image, the result is likely to fail. The other issue is the accumulated error from the 2D pose estimation result. The proposed 3D pose estimation network utilizes the 2D pose estimation results as input to predict the 3D pose, which results in accumulated error. Therefore, a new network or method for the multiple-machine situation and the 3D pose direct training can be designed in the future work.

The accuracy of the proposed 2D pose estimation method is 40.64 mm in the laboratory dataset, which is acceptable for the object detection and tracking and the proximity detection application discussed in Table 2. On the other hand, the accuracy of the proposed 3D pose estimation method is 144.65 mm, which is not adequate for preventing utility strikes, grade control, and autonomous excavation applications, even though it may be acceptable in proximity detection applications.

The camera distance is important for pose estimation on

construction sites. The scope of some existing human pose estimation method is not suitable for the articulated construction robot due to short camera distance. The typical excavator working range is within 6.1 m, according to Lundeen et al. [12]. Thus, the performance of the articulated construction robot pose estimation should be evaluated over 6.1 m for the camera distance. The camera distance in the evaluation of the proposed method is 10 m.

The proposed pose estimation method has three limitations. First, the network trained on the laboratory dataset was unable to achieve high performance when applied to an excavator operating in the field. The background and the light conditions of the laboratory dataset do not have a wide variety since they were collected in the same indoor environment, compared to actual construction sites where such conditions may vary. Second, the latency of the proposed method is affected by the hardware specifications and the complexity of the network architecture, which would need extra cost for the advanced hardware in order to achieve the great performance. Third, the system assumes the consistency and quality of the source video stream, which is not always available in real practice, especially on hazardous and unstructured construction sites. Further research of the data consistency on construction sites needs to be conducted to explore this issue. Fourth, the proposed 3D pose estimation method is trained and evaluated on the laboratory dataset due to lack of the 3D ground truth data for the real site dataset. Future work on augmenting the real site dataset with ground truth data from onboard sensors of the excavator, or exploring new network to train without ground truth data need to be conducted.

8. Conclusions and future work

In this research, vision-based marker-less 2D and 3D pose estimation methods for articulated construction robots were proposed, in which an excavator was used as the experimental machine test-bed. The excavator boom, stick, and bucket joint positions are estimated with both 2D and 3D coordinates. A state-of-the-art human pose estimation deep convolutional network, i.e., the stacked hourglass network, was adapted and modified for the application. The network model was trained on an excavator dataset, which was collected and annotated with a KUKA robot arm representing an excavator and from real construction sites with working excavators. The sensor-based pose estimation method was also implemented to evaluate the performance of the proposed network. The results showed that the proposed network could estimate the boom and stick joints but had higher estimation error for the bucket location due to typically encountered occlusion issues.

Moreover, the accumulated error in the 3D pose estimation resulting from the 2D predicted pose input needs to be resolved as well. Therefore, in proposed future work, additional training image data with higher variety will be collected. A further modification of the proposed network will also be explored to adapt to the multiple-machine situation and address the accumulated error issues. Finally, the data consistency on construction sites will also be considered and surveyed.

Acknowledgments

The work presented in this paper was supported financially by a United States National Science Foundation National Science Foundation Award (No. IIS-1734266, *'Scene Understanding and Predictive Monitoring for Safe Human-Robot Collaboration in Unstructured and Dynamic Construction Environments'*). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Z. Zhou, Y.M. Goh, Q. Li, Overview and analysis of safety management studies in the construction industry, Saf. Sci. 72 (2015) 337–350, https://doi.org/10.1016/j.ssci.2014.10.006.

[2] C.-J. Liang, S.-C. Kang, M.-H. Lee, RAS: a robotic assembly system for steel structure erection and assembly, Int. J. Intell. Robot. Appl. 1 (2017) 459–476, https://doi.org/10.1007/s41315-017-0030-x.

[3] BLS, An analysis of fatal occupational injuries at road construction sites, 2003–2010, Mon. Labor Rev. (2013). https://stats.bls.gov/opub/mlr/2013/article/pdf/an-analysis-of-fatal-occupational-injuries-at-road-construction-sites-2003-2010.pdf (accessed February 27, 2019).

[4] CPWR (Ed.), The Construction Chart Book: The U.S. Construction Industry and its Workers, 4th ed, CPWR - The Center for Construction Research and Training, Silver Spring, MD, 2008(ISBN:978-0-9802115-0-4).

[5] W.-H. Hung, C.-W. Liu, C.-J. Liang, S.-C. Kang, Strategies to accelerate the computation of erection paths for construction cranes, Autom. Constr. 62 (2016) 1–13, https://doi.org/10.1016/j.autcon.2015.10.008.

[6] J. Teizer, B.S. Allread, U. Mantripragada, Automating the blind spot measurement of construction equipment, Autom. Constr. 19 (2010) 491–501, https://doi.org/10.1016/j.autcon.2009.12.012.

[7] J. Hinze, R. Godfrey, An evaluation of safety performance measures for construction projects, J. Constr. Res. 04 (2003) 5–15, https://doi.org/10.1142/S160994510300025X.

[8] R.E. Levitt, N.M. Samelson, Construction Safety Management, John Wiley & Sons, 978-0-471-59933-3, 1993.

[9] S. Talmaki, V.R. Kamat, Real-time hybrid virtuality for prevention of excavation related utility strikes, J. Comput. Civ. Eng. 28 (2014) 04014001, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000269.

[10] A.H. Behzadan, V.R. Kamat, Interactive augmented reality visualization for improved damage prevention and maintenance of underground infrastructure, Proceedings of the Construction Research Congress (CRC), ASCE, Seattle, WA, USA, 2009, pp. 1214–1222, , https://doi.org/10.1061/41020(339)123.

[11] Common Ground Alliance, New common ground alliance dirt report estimates that damage to buried utilities cost society at least $1.5 billion last year, http://commongroundalliance.com/media-reports/press-releases/new-common-ground-alliance-dirt-report-estimates-damage-buried, (2017) , Accessed date: 15 August 2018.

[12] K.M. Lundeen, S. Dong, N. Fredricks, M. Akula, J. Seo, V.R. Kamat, Optical marker-based end effector pose estimation for articulated excavators, Autom. Constr. 65 (2016) 51–64, https://doi.org/10.1016/j.autcon.2016.02.003.

[13] S. Li, H. Cai, V.R. Kamat, Uncertainty-aware geospatial system for mapping and visualizing underground utilities, Autom. Constr. 53 (2015) 105–119, https://doi.org/10.1016/j.autcon.2015.03.011.

[14] E. Rezazadeh Azar, B. McCabe, Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos, Autom. Constr. 24 (2012) 194–202, https://doi.org/10.1016/j.autcon.2012.03.003.

[15] M.M. Soltani, Z. Zhu, A. Hammad, Framework for location data fusion and pose estimation of excavators using stereo vision, J. Comput. Civ. Eng. 32 (2018) 04018045, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000783.

[16] F. Vahdatikhaki, A. Hammad, H. Siddiqui, Optimization-based excavator pose estimation using real-time location systems, Autom. Constr. 56 (2015) 76–92, https://doi.org/10.1016/j.autcon.2015.03.006.

[17] V.R. Kamat, A.H. Behzadan, GPS and 3DOF tracking for georeferenced registration of construction graphics in outdoor augmented reality, Proceedings of the EG-ICE International Workshop on Intelligent Computing in Engineering and Architecture, Springer, Berlin, Heidelberg, Ascona, Switzerland, 2006, pp. 368–375, , https://doi.org/10.1007/11888598_34.

[18] A.H. Behzadan, V.R. Kamat, Animation of construction activities in outdoor augmented reality, Proceedings of the Joint International Conference on Computing and Decision Making in Civil and Building Engineering (ICCCBE), Montréal, Canada, 2006, pp. 1135–1143 http://pathfinder.engin.umich.edu/documents/Behzadan&Kamat.ICCCBEXI.2006.pdf , Accessed date: 2 April 2019.

[19] A.H. Behzadan, V.R. Kamat, Integrated information modeling and visual simulation of engineering operations using dynamic augmented reality scene graphs, Electron. J. Inf. Technol. Constr. 16 (2011) 259–278 http://www.itcon.org/paper/2011/17 , Accessed date: 28 March 2019.

[20] C. Feng, S. Dong, K.M. Lundeen, Y. Xiao, V.R. Kamat, Vision-based articulated machine pose estimation for excavation monitoring and guidance, Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC), IAARC, Oulu, Finland, 2015, , https://doi.org/10.22260/ISARC2015/0029.

[21] E. Rezazadeh Azar, S. Dickinson, B. McCabe, Server-customer interaction tracker: computer vision–based system to estimate dirt-loading cycles, J. Constr. Eng. Manag. 139 (2013) 785–794, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000652.

[22] A. Montaser, O. Moselhi, RFID+ for tracking earthmoving operations, Proceedings of the Construction Research Congress (CRC), ASCE, West Lafayette, IN, USA, 2012, pp. 1011–1020, , https://doi.org/10.1061/9780784412329.102.

[23] M. Ibrahim, O. Moselhi, Automated productivity assessment of earthmoving operations, Electron. J. Inf. Technol. Constr. 19 (2014) 169–184 http://www.itcon.org/paper/2014/9 , Accessed date: 26 February 2019.

[24] J. Gong, C.H. Caldas, An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations, Autom. Constr. 20 (2011) 1211–1226, https://doi.org/10.1016/j.autcon.2011.05.005.

[25] C. Yuan, S. Li, H. Cai, Vision-based excavator detection and tracking using hybrid kinematic shapes and key nodes, J. Comput. Civ. Eng. 31 (2017) 04016038, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000602.

[26] J. Kim, S. You, S. Lee, V.R. Kamat, L.P. Robert, Evaluation of human robot collaboration in masonry work using immersive virtual environments, Proceedings of

the International Conference on Construction Applications of Virtual Reality (CONVR), Banff, Canada, 2015, pp. 132–141 http://pathfinder.engin.umich.edu/documents/KimEtAl.CONVR.2015.pdf , Accessed date: 27 February 2019.
[27] T. Salmi, J.M. Ahola, T. Heikkilä, P. Kilpeläinen, T. Malm, Human-robot collaboration and sensor-based robots in industrial applications and construction, in: H. Bier (Ed.), Robotic Building, Springer International Publishing, Cham, 2018, pp. 25–52, , https://doi.org/10.1007/978-3-319-70866-9_2.
[28] T. Salmi, I. Marstio, T. Malm, J. Montonen, Advanced safety solutions for human-robot-cooperation, Proceedings of the International Symposium on Robotics (ISR), Munich, Germany, 2016, pp. 610–615 https://cris.vtt.fi/en/publications/advanced-safety-solutions-for-human-robot-cooperation , Accessed date: 9 September 2018.
[29] P.D. Groves, Shadow matching: a new GNSS positioning technique for urban canyons, J. Navig. 64 (2011) 417–430, https://doi.org/10.1017/S0373463311000087.
[30] C. Feng, Y. Xiao, A. Willette, W. McGee, V.R. Kamat, Towards autonomous robotic in-situ assembly on unstructured construction sites using monocular vision, Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC), IAARC, Sydney, Australia, 2014, pp. 163–170, , https://doi.org/10.22260/ISARC2014/0022.
[31] S. Talmaki, V.R. Kamat, Multi-sensor monitoring for real-time 3D visualization of construction equipment, Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC), IAARC, Montréal, Canada, 2013, pp. 27–43, , https://doi.org/10.22260/ISARC2013/0004.
[32] F.A. Bender, S. Göltz, T. Bräunl, O. Sawodny, Modeling and offset-free model predictive control of a hydraulic mini excavator, IEEE Trans. Autom. Sci. Eng. 14 (2017) 1682–1694, https://doi.org/10.1109/TASE.2017.2700407.
[33] Z. Péntek, T. Hiller, T. Liewald, B. Kuhlmann, A. Czmerk, IMU-based mounting parameter estimation on construction vehicles, Proceedings of the DGON Inertial Sensors and Systems (ISS), IEEE, Karlsruhe, Germany, 2017, pp. 1–14, , https://doi.org/10.1109/InertialSensors.2017.8171504.
[34] H. Kim, C.R. Ahn, D. Engelhaupt, S. Lee, Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement, Autom. Constr. 87 (2018) 225–234, https://doi.org/10.1016/j.autcon.2017.12.014.
[35] C.R. Ahn, S. Lee, F. Peña-Mora, Application of low-cost accelerometers for measuring the operational efficiency of a construction equipment fleet, J. Comput. Civ. Eng. 29 (2015) 04014042, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000337.
[36] J. Park, J. Chen, Y.K. Cho, Self-corrective knowledge-based hybrid tracking system using BIM and multimodal sensors, Adv. Eng. Inform. 32 (2017) 126–138, https://doi.org/10.1016/j.aei.2017.02.001.
[37] Z. Aziz, C.J. Anumba, D. Ruikar, P.M. Carrillo, N.M. Bouchlaghem, Context aware information delivery for on-site construction operations, Proceedings of the CIB-W78 International Conference on Information Technology in Construction, CIB Publication, Dresden, Germany, 2005, pp. 321–332 (ISBN:3-86005-478-3).
[38] C. Rohrig, F. Kiinemund, Mobile robot localization using WLAN signal strengths, Proceedings of the IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IEEE, Dortmund, Germany, 2007, pp. 704–709, , https://doi.org/10.1109/IDAACS.2007.4488514.
[39] B.-W. Jo, Y.-S. Lee, J.-H. Kim, D.-K. Kim, P.-H. Choi, Proximity warning and excavator control system for prevention of collision accidents, Sustainability 9 (2017) 1488, https://doi.org/10.3390/su9081488.
[40] H.M. Khoury, V.R. Kamat, Evaluation of position tracking technologies for user localization in indoor construction environments, Autom. Constr. 18 (2009) 444–457, https://doi.org/10.1016/j.autcon.2008.10.011.
[41] J. Teizer, M. Venugopal, A. Walia, Ultrawideband for automated real-time three-dimensional location sensing for workforce, equipment, and material positioning and tracking, Transp. Res. Rec. 2081 (2008) 56–64, https://doi.org/10.3141/2081-06.
[42] C. Zhang, A. Hammad, S. Rodriguez, Crane pose estimation using UWB real-time location system, J. Comput. Civ. Eng. 26 (2012) 625–637, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000172.
[43] J. Chai, C. Wu, C. Zhao, H.-L. Chi, X. Wang, B.W.-K. Ling, K.L. Teo, Reference tag supported RFID tracking using robust support vector regression and Kalman filter, Adv. Eng. Inform. 32 (2017) 1–10, https://doi.org/10.1016/j.aei.2016.11.002.
[44] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Adv. Eng. Inform. 29 (2015) 239–251, https://doi.org/10.1016/j.aei.2015.02.001.
[45] J. Chen, Y. Fang, Y.K. Cho, C. Kim, Principal axes descriptor for automated construction-equipment classification from point clouds, J. Comput. Civ. Eng. 31 (2017) 04016058, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000628.
[46] E. Rezazadeh Azar, B. McCabe, Automated visual recognition of dump trucks in construction videos, J. Comput. Civ. Eng. 26 (2012) 769–781, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000179.
[47] M.M. Soltani, Z. Zhu, A. Hammad, Automated annotation for visual recognition of construction resources using synthetic images, Autom. Constr. 62 (2016) 14–23, https://doi.org/10.1016/j.autcon.2015.10.002.
[48] C.-J. Liang, V.R. Kamat, C.C. Menassa, Real-time construction site layout and equipment monitoring, Proceedings of the Construction Research Congress, New Orleans, LA, 2018, pp. 64–74, , https://doi.org/10.1061/9780784481264.007.
[49] M.M. Soltani, Z. Zhu, A. Hammad, Skeleton estimation of excavator by detecting its parts, Autom. Constr. 82 (2017) 1–15, https://doi.org/10.1016/j.autcon.2017.06.023.
[50] C. Feng, V.R. Kamat, H. Cai, Camera marker networks for articulated machine pose estimation, Autom. Constr. 96 (2018) 148–160, https://doi.org/10.1016/j.autcon.2018.09.004.
[51] E. Rezazadeh Azar, C. Feng, V.R. Kamat, Feasibility of in-plane articulation monitoring of excavator arm using planar marker tracking, Electron. J. Inf. Technol. Constr. 20 (2015) 213–229 http://itcon.org/paper/2015/15 , Accessed date: 13 February 2017.
[52] W. Yang, X. Zhang, H. Ma, G.-M. Zhang, Infrared LEDs-based pose estimation with underground camera model for boom-type roadheader in coal mining, IEEE Access 7 (2019) 33698–33712, https://doi.org/10.1109/ACCESS.2019.2904097.
[53] C.-J. Liang, Y.-Y. Yang, Y.-S. Lin, S.-C. Kang, P.-C. Lin, Y.-C. Chen, Botbeep - an affordable warning device for wheelchair rearward safety, Proceedings of the IEEE International Conference on Orange Technologies (ICOT), IEEE, Tainan, Taiwan, 2013, pp. 159–163, , https://doi.org/10.1109/ICOT.2013.6521182.
[54] C. Feng, V.R. Kamat, Plane registration leveraged by global constraints for context-aware AEC applications, Comput. Aided Civ. Inf. Eng. 28 (2013) 325–343, https://doi.org/10.1111/j.1467-8667.2012.00795.x.
[55] L. Xu, V.R. Kamat, C.C. Menassa, Automatic extraction of 1D barcodes from video scans for drone-assisted inventory management in warehousing applications, Int J Log Res Appl (2017) 1–16, https://doi.org/10.1080/13675567.2017.1393505.
[56] C. Feng, Y. Xiao, A. Willette, W. McGee, V.R. Kamat, Vision guided autonomous robotic assembly and as-built scanning on unstructured construction sites, Autom. Constr. 59 (2015) 128–138, https://doi.org/10.1016/j.autcon.2015.06.002.
[57] B.R.K. Mantha, C.C. Menassa, V.R. Kamat, Robotic data collection and simulation for evaluation of building retrofit performance, Autom. Constr. 92 (2018) 88–102, https://doi.org/10.1016/j.autcon.2018.03.026.
[58] J. Seo, R. Starbuck, S. Han, S. Lee, T.J. Armstrong, Motion data-driven biomechanical analysis during construction tasks on sites, J. Comput. Civ. Eng. 29 (2015) B4014005, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000400.
[59] S. Han, S. Lee, F. Peña-Mora, Comparative study of motion features for similarity-based modeling and classification of unsafe actions in construction, J. Comput. Civ. Eng. 28 (2014) A4014005, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000339.
[60] S. Han, S. Lee, A vision-based motion capture and recognition framework for behavior-based safety management, Autom. Constr. 35 (2013) 131–141, https://doi.org/10.1016/j.autcon.2013.05.001.
[61] S. Han, S. Lee, F. Peña-Mora, Vision-based detection of unsafe actions of a construction worker: case study of ladder climbing, J. Comput. Civ. Eng. 27 (2013) 635–644, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000279.
[62] K.M. Lundeen, V.R. Kamat, C.C. Menassa, W. McGee, Scene understanding for adaptive manipulation in robotized construction work, Autom. Constr. 82 (2017) 16–30, https://doi.org/10.1016/j.autcon.2017.06.022.
[63] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D human pose estimation: new benchmark and state of the art analysis, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Columbus, OH, USA, 2014, pp. 3686–3693, , https://doi.org/10.1109/CVPR.2014.471.
[64] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, B. Schiele, DeepCut: joint subset partition and labeling for multi person pose estimation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 4929–4937, , https://doi.org/10.1109/CVPR.2016.533.
[65] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, DeeperCut: a deeper, stronger, and faster multi-person pose estimation model, Proceedings of the European Conference on Computer Vision (ECCV), Springer, Cham, Amsterdam, Netherlands, 2016, pp. 34–50, , https://doi.org/10.1007/978-3-319-46466-4_3.
[66] D.C. Luvizon, D. Picard, H. Tabia, 2D/3D pose estimation and action recognition using multitask deep learning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, USA, 2018, pp. 5137–5146, , https://doi.org/10.1109/CVPR.2018.00539.
[67] A. Bulat, G. Tzimiropoulos, Human pose estimation via convolutional part heatmap regression, Proceedings of the European Conference on Computer Vision (ECCV), Springer International Publishing, Amsterdam, Netherlands, 2016, pp. 717–732, , https://doi.org/10.1007/978-3-319-46478-7_44.
[68] A. Toshev, C. Szegedy, DeepPose: human pose estimation via deep neural networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Columbus, OH, USA, 2014, pp. 1653–1660, , https://doi.org/10.1109/CVPR.2014.214.
[69] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, Proceedings of the European Conference on Computer Vision (ECCV), Springer, Cham, Amsterdam, Netherlands, 2016, pp. 483–499, , https://doi.org/10.1007/978-3-319-46484-8_29.
[70] Y. Chen, C. Shen, X.-S. Wei, L. Liu, J. Yang, Adversarial PoseNet: a structure-aware convolutional network for human pose estimation, Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, 2017, pp. 1212–1221, , https://doi.org/10.1109/ICCV.2017.137.
[71] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, 2017, pp. 1281–1290, , https://doi.org/10.1109/ICCV.2017.144.
[72] X. Chu, W. Yang, W. Ouyang, C. Ma, A.L. Yuille, X. Wang, Multi-context attention for human pose estimation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA, 2017, pp. 1831–1840, , https://doi.org/10.1109/CVPR.2017.601.
[73] G. Pavlakos, X. Zhou, K.G. Derpanis, K. Daniilidis, Coarse-to-fine volumetric prediction for single-image 3D human pose, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA, 2017, pp. 7025–7034, , https://doi.org/10.1109/CVPR.2017.139.
[74] X. Zhou, Q. Huang, X. Sun, X. Xue, Y. Wei, Towards 3D human pose estimation in the wild: a weakly-supervised approach, Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, 2017, pp. 398–407, ,

https://doi.org/10.1109/ICCV.2017.51.
[75] J. Martinez, R. Hossain, J. Romero, J.J. Little, A simple yet effective baseline for 3d human pose estimation, Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, 2017, pp. 2659–2668, , https://doi.org/10.1109/ICCV.2017.288.
[76] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1325–1339, https://doi.org/10.1109/TPAMI.2013.248.
[77] Trimble, Grade control for compact machines, https://construction.trimble.com/products-and-solutions/grade-control-compact-machines, (2018) , Accessed date: 27 September 2018.
[78] John Deere US, Grade control: construction technology solutions, https://www.deere.com/en/construction/construction-technology/grade-control/, (2018) , Accessed date: 27 September 2018.
[79] R. Maalek, F. Sadeghpour, Accuracy assessment of Ultra-Wide Band technology in tracking static resources in indoor construction scenarios, Autom. Constr. 30 (2013) 170–183, https://doi.org/10.1016/j.autcon.2012.10.005.
[80] M. Memarzadeh, M. Golparvar-Fard, J.C. Niebles, Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors, Autom. Constr. 32 (2013) 24–37, https://doi.org/10.1016/j.autcon.2012.12.002.
[81] J. Wang, S.N. Razavi, Low false alarm rate model for unsafe-proximity detection in construction, J. Comput. Civ. Eng. 30 (2016) 04015005, , https://doi.org/10.1061/(ASCE)CP.1943-5487.0000470.
[82] D. Kim, K. Yin, M. Liu, S. Lee, V.R. Kamat, Feasibility of a drone-based on-site proximity detection in an outdoor construction site, Proceedings of the ASCE International Workshop on Computing in Civil Engineering (IWCCE), ASCE, Seattle, WA, USA, 2017, pp. 392–400, , https://doi.org/10.1061/9780784480847.049.
[83] G.J. Maeda, D.C. Rye, S.P.N. Singh, Iterative autonomous excavation, Proceedings of the International Conference on Field and Service Robotics (FSR), Matsushima, Japan, 2012, pp. 369–382, , https://doi.org/10.1007/978-3-642-40686-7_25.
[84] H. Shao, H. Yamamoto, Y. Sakaida, T. Yamaguchi, Y. Yanagisawa, A. Nozue, Automatic excavation planning of hydraulic excavator, Proceedings of the International Conference on Intelligent Robotics and Applications, Springer, Berlin, Heidelberg, Wuhan, China, 2008, pp. 1201–1211, , https://doi.org/10.1007/978-3-540-88518-4_128.
[85] A. Newell, Z. Huang, J. Deng, Associative embedding: end-to-end learning for joint detection and grouping, Advances in Neural Information Processing Systems, NIPS, Long Beach, CA, USA, 2017, pp. 2274–2284 http://papers.nips.cc/paper/6822-associative-embedding-end-to-end-learning-for-joint-detection-and-grouping , Accessed date: 27 February 2019.
[86] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 2015, pp. 448–456 http://proceedings.mlr.press/v37/ioffe15.html , Accessed date: 29 September 2018.
[87] KUKA Robotics Corporation, KR QUANTEC pro, https://www.kuka.com/en-us/products/robotics-systems/industrial-robots/kr-quantec-pro, (2018) , Accessed date: 7 October 2018.
[88] FLIR, FLIR USB 3.1, Gigabit Ethernet and FireWire Machine Vision Cameras, https://www.ptgrey.com/, (2018) , Accessed date: 7 October 2018.
[89] L. Perez, J. Wang, The Effectiveness of Data Augmentation in Image Classification Using Deep Learning, ArXiv:1712.04621 [Cs], 2017, http://arxiv.org/abs/1712.04621 , Accessed date: 26 February 2019.
[90] B. Sapp, B. Taskar, MODEC: multimodal decomposable models for human pose estimation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Portland, OR, USA, 2013, pp. 3674–3681, , https://doi.org/10.1109/CVPR.2013.471.
[91] Xsens, MTw Awinda, Xsens 3D motion tracking, https://www.xsens.com/products/mtw-awinda/, (2018) , Accessed date: 7 October 2018.
[92] J.M. Wong, V. Kee, T. Le, S. Wagner, G.-L. Mariottini, A. Schneider, L. Hamilton, R. Chipalkatty, M. Hebert, D.M.S. Johnson, J. Wu, B. Zhou, A. Torralba, SegICP: integrated deep semantic segmentation and pose estimation, Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Vancouver, BC, Canada, 2017, pp. 5784–5789, , https://doi.org/10.1109/IROS.2017.8206470.
[93] J. Wu, B. Zhou, R. Russell, V. Kee, S. Wagner, M. Hebert, A. Torralba, D.M.S. Johnson, Real-time object pose estimation with pose interpreter networks, Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 2018, pp. 6798–6805, , https://doi.org/10.1109/IROS.2018.8593662.
[94] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, C. Theobalt, Vnect: real-time 3D human pose estimation with a single rgb camera, ACM Trans. Graph. 36 (2017) 1–14, https://doi.org/10.1145/3072959.3073596.
[95] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA, 2017, pp. 7291–7299 http://openaccess.thecvf.com/content_cvpr_2017/html/Cao_Realtime_Multi-Person_2D_CVPR_2017_paper.html , Accessed date: 3 April 2019.