Fast Dataset Collection Approach for Articulated Equipment Pose Estimation

Ci-Jyun Liang[1]; Kurt M. Lundeen[2]; Wes McGee[3]; Carol C. Menassa, Ph.D.[4];
SangHyun Lee, Ph.D.[5]; and Vineet R. Kamat, Ph.D.[6]

[1]Ph.D. Candidate, Laboratory for Interactive Visualization in Engineering, Dept. of Civil and Environmental Engineering, Univ. of Michigan, 2350 Hayward St., 2105 G. G. Brown Building, Ann Arbor, MI 48109-2125. E-mail: cjliang@umich.edu
[2]Ph.D. Candidate, Laboratory for Interactive Visualization in Engineering, Dept. of Civil and Environmental Engineering, Univ. of Michigan, 2350 Hayward St., 2105 G. G. Brown Building, Ann Arbor, MI 48109-2125. E-mail: klundeen@umich.edu
[3]Fabrication Lab, Taubman College of Architecture and Urban Planning, Univ. of Michigan, 2000 Bonisteel Blvd., Ann Arbor, MI 48109-2069. E-mail: wesmcgee@umich.edu
[4]Sustainable and Intelligent Civil Infrastructure Systems Laboratory, Dept. of Civil and Environmental Engineering, Univ. of Michigan, 2350 Hayward St., 2140 G. G. Brown Building, Ann Arbor, MI 48109-2125. E-mail: menassa@umich.edu
[5]Dynamic Project Management Group, Dept. of Civil and Environmental Engineering, Univ. of Michigan, 2350 Hayward St., 2012 G. G. Brown Building, Ann Arbor, MI 48109-2125. E-mail: shdpm@umich.edu
[6]Laboratory for Interactive Visualization in Engineering, Dept. of Civil and Environmental Engineering, Univ. of Michigan, 2350 Hayward St., 2008 G. G. Brown Building, Ann Arbor, MI 48109-2125. E-mail: vkamat@umich.edu

ABSTRACT

Struck-by accidents are potential safety concerns on construction sites and require a robust machine pose estimation. The development of deep learning methods has enhanced the human pose estimation that can be adapted for articulated machines. These methods require abundant dataset for training, which is challenging and time-consuming to obtain on-site. This paper proposes a fast data collection approach to build the dataset for excavator pose estimation. It uses two industrial robot arms as the excavator and the camera monopod to collect different excavator pose data. The 3D annotation can be obtained from the robot's embedded encoders. The 2D pose is annotated manually. For evaluation, 2,500 pose images were collected and trained with the stacked hourglass network. The results showed that the dataset is suitable for the excavator pose estimation network training in a controlled environment, which leads to the potential of the dataset augmenting with real construction site images.

INTRODUCTION

The prospect of human-robot collaboration (HRC) on construction sites raises safety concerns (Liang et al. 2018b; You et al. 2018). Unlike HRC in typical industrial 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 ISO standards, 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 (Salmi et al. 2018). 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. However, the protective distance 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 (Salmi et al. 2018). This highlights the need for developing an effective on-site pose estimation system for articulated construction equipment and human workers, as shown in Figure 1.

Vision-based methods can extract object's pose directly from input data with a marker (marker-based) or without marker (marker-less) (Liang et al. 2018b). The marker-based method identifies all the markers mounted on equipment and estimates the pose by their geometric relations or marker network (Feng et al. 2018; Liang et al. 2017; Lundeen et al. 2016; Rezazadeh Azar et al. 2015), whereas the marker-less method directly extracts features and estimates the pose from them (Liang et al. 2018a; Soltani et al. 2018). 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 (Han et al. 2013, 2014; Han and Lee 2013; Seo et al. 2015). Feature descriptor is the first type of marker-less pose estimation method (Chen et al. 2017; Lundeen et al. 2017; Rezazadeh Azar et al. 2013). The recently emerging Convolutional Neural Networks (CNN) is another type of pose estimation method (Andriluka et al. 2014), 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 (Newell et al. 2016), which estimate the human pose in 2D pixel-wise coordinates, as shown in Figure 1. This is due to the lack of 3D ground truth posture data (Martinez et al. 2017). 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 (Ionescu et al. 2014), which is difficult to employ for construction equipment in an outdoor environment.

In this study, a fast dataset collection approach for articulated equipment pose estimation is developed and evaluated. This approach collects images from a factory environment with a robotic manipulator and from real construction sites. Both 2D and 3D data are annotated. The performance of the dataset is validated by a state-of-the-art 2D human pose estimation network (Newell et al. 2016) and a 3D human pose estimation baseline network (Martinez et al. 2017), and compared with the IMU sensor pose estimation method.

DATA COLLECTION APPROACH

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 2D and 3D networks are trained by the training group and then evaluated by the testing group.

Dataset Collection Setup: For the dataset collection setup, a KUKA seven degrees-of-freedom (DOF) robot arm (KUKA KR120) was used to simulate the excavator, and the images of the robot arm with different poses were captured. Figure 2 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. 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 $\left(X, Y, Z, A, B, C\right)$ and angles of all joints $\left(A_1, A_2, A_3, A_4, A_5, A_6\right)$.

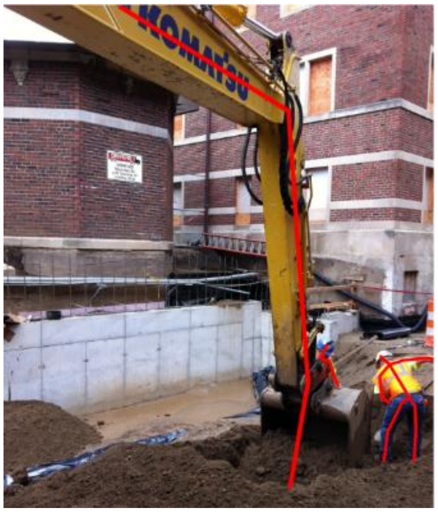

Figure 1. Illustration of the 2D on-site pose estimation system on a video frame for both articulated equipment and human workers. Red lines are the estimated pose.

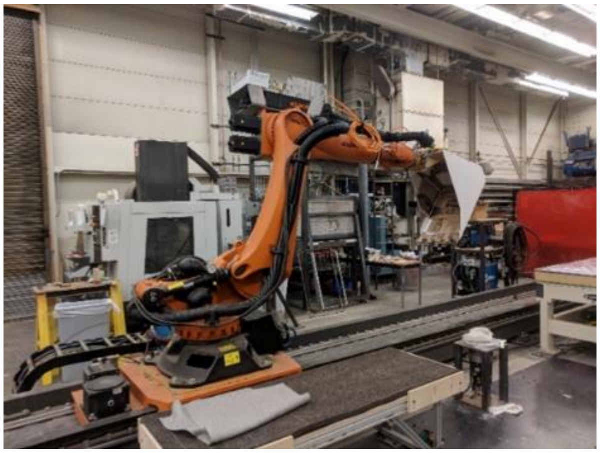

Figure 2. The simulated robotic excavator - robot arm mounted with an excavator bucket.

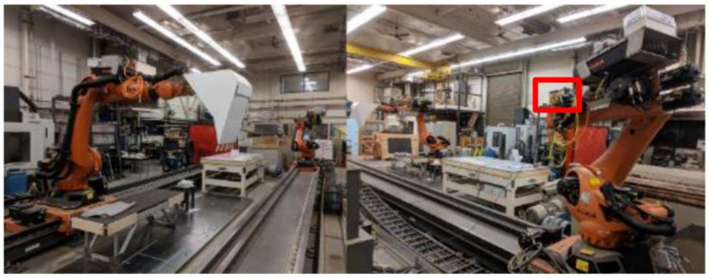

Figure 3. The camera mounted on the second robot arm to capture the images (red square).

In order to collect the images of the simulated excavator, a Point Grey camera was used in the process. The camera was mounted on a second KUKA robot arm in the laboratory, as shown in Figure 3. 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 2,500 images were collected; 2,000 of them were used as training images and 500 of them were used as testing images. Figure 4 shows a set of the collected images from the dataset. The size of each image is 2048x2048 pixels.

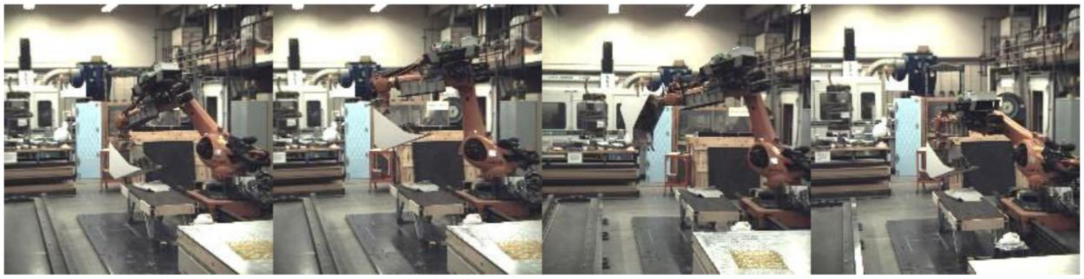

Figure 4. A set of the captured images for the excavator dataset with different camera location and orientation, and excavator pose.

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 the 2D pose (Andriluka et al. 2014) and Human3.6M for the 3D pose (Ionescu et al. 2014). In the 2D pose annotation, excavator joint locations were annotated in the pixel-wise coordinate. 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 distance 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. Figure 5 shows an example of an annotated image.

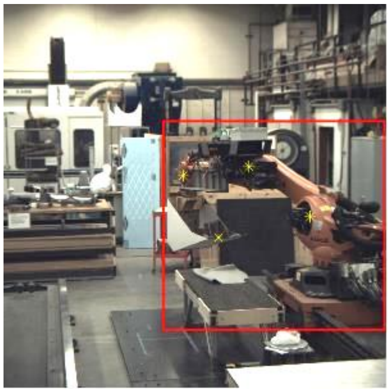

Figure 5. An example of the annotated image. Stars represent the joint locations and the rectangle represents the bounding box.

The 3D ground truth data was acquired by the robot arm's built-in encoders and the PLC. 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.

Sensor-based Pose Estimation: For evaluating the excavator dataset, 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. These sensors were placed on the axis of each joint so that they can measure the correct angle when the robot 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 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.

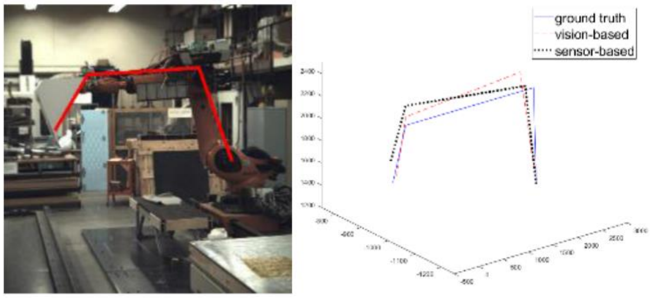

Figure 6. Results of the excavator 3D pose estimation. The left image is the result from 2D pose estimation and the right image is the 3D result.

Table 1. Results of the average Euclidean distance (mm) between the predicted and the ground truth joint location.

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

EXPERIMENT RESULTS

The proposed pose estimation dataset was evaluated by comparing the estimated results and the ground truth, as shown in Figure 6. The left image was the 2D result and the right image was the 3D 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 1. The average Euclidean distance between the 3D vision-based method and ground truth is 144.65 mm, and between the sensor-based 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 reason was that the camera coordinates preprocessing mentioned in (Martinez et al. 2017) was not applied to the ground truth data because the camera matrix was not determined in the laboratory dataset. In addition, the occlusion issue also affected the prediction result.

CONCLUSION

In this research, a fast dataset collection approach for 2D and 3D articulated construction robots pose estimation methods were proposed. A KUKA robot arm was utilized to represent an excavator in a factory and a camera on the second robot arm was used to capture the image. The 3D pose was acquired from robot arm sensors and the 2D pose was annotated manually. A 3D pose estimation network was evaluated on the dataset. The sensor-based pose estimation method was also implemented to compare the performance. The results showed that the dataset collected by the proposed approach could estimate excavator's joints but had higher estimation error.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support for this research received from the US National Science Foundation (NSF) via grant 1734266. Any opinions and findings in this paper are those of the authors and do not necessarily represent those of the NSF.

REFERENCES

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). “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, 3686–3693.

Chen, J., Fang, Y., Cho, Y. K., and Kim, C. (2017). “Principal axes descriptor for automated construction-equipment classification from point clouds.” *Journal of Computing in Civil Engineering*, 31(2), 04016058.

Feng, C., Kamat, V. R., and Cai, H. (2018). “Camera marker networks for articulated machine pose estimation.” *Automation in Construction*, 96, 148–160.

Han, S., and Lee, S. (2013). “A vision-based motion capture and recognition framework for behavior-based safety management.” *Automation in Construction*, 35, 131–141.

Han, S., Lee, S., and Peña-Mora, F. (2013). “Vision-based detection of unsafe actions of a construction worker: case study of ladder climbing.” *Journal of Computing in Civil Engineering*, 27(6), 635–644.

Han, S., Lee, S., and Peña-Mora, F. (2014). “Comparative study of motion features for similarity-based modeling and classification of unsafe actions in construction.” *Journal of Computing in Civil Engineering*, 28(5), A4014005.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). “Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 1325–1339.

Liang, C.-J., Kamat, V. R., and Menassa, C. C. (2018a). “Real-time construction site layout and equipment monitoring.” *Proceedings of the 2018 Construction Research Congress*, New Orleans, LA, 64–74.

Liang, C.-J., Kang, S.-C., and Lee, M.-H. (2017). “RAS: a robotic assembly system for steel structure erection and assembly.” *International Journal of Intelligent Robotics and Applications*, 1(4), 459–476.

Liang, C.-J., Lundeen, K. M., McGee, W., Menassa, C. C., Lee, S., and Kamat, V. R. (2018b).

"Stacked hourglass networks for markerless pose estimation of articulated construction robots." *Proceedings of the 35th International Symposium on Automation and Robotics in Construction*, Berlin, Germany.

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

Lundeen, K. M., Kamat, V. R., Menassa, C. C., and McGee, W. (2017). "Scene understanding for adaptive manipulation in robotized construction work." *Automation in Construction*, 82, 16–30.

Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). "A simple yet effective baseline for 3d human pose estimation." *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, IEEE, Venice, Italy, 2659–2668.

Newell, A., Yang, K., and Deng, J. (2016). "Stacked hourglass networks for human pose estimation." *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, Springer, Cham, Amsterdam, Netherlands, 483–499.

Rezazadeh Azar, E., Dickinson, S., and McCabe, B. (2013). "Server-customer interaction tracker: computer vision–based system to estimate dirt-loading cycles." *Journal of Construction Engineering and Management*, 139(7), 785–794.

Rezazadeh Azar, E., Feng, C., and Kamat, V. R. (2015). "Feasibility of in-plane articulation monitoring of excavator arm using planar marker tracking." *Journal of Information Technology in Construction (ITcon)*, 20(15), 213–229.

Salmi, T., Ahola, J. M., Heikkilä, T., Kilpeläinen, P., and Malm, T. (2018). "Human-robot collaboration and sensor-based robots in industrial applications and construction." *Robotic Building*, Springer Series in Adaptive Environments, H. Bier, ed., Springer International Publishing, Cham, 25–52.

Seo, J., Starbuck, R., Han, S., Lee, S., and Armstrong, T. J. (2015). "Motion data-driven biomechanical analysis during construction tasks on sites." *Journal of Computing in Civil Engineering*, 29(4), B4014005.

Soltani, M. M., Zhu, Z., and Hammad, A. (2018). "Framework for location data fusion and pose estimation of excavators using stereo vision." *Journal of Computing in Civil Engineering*, 32(6), 04018045.

You, S., Kim, J.-H., Lee, S., Kamat, V., and Robert, L. P. (2018). "Enhancing perceived safety in human–robot collaborative construction using immersive virtual environments." *Automation in Construction*, 96, 161–170.