Digital twinning construction objects: Lessons learned from pose estimation methods

Fan Xue[1], Hongling Guo[2], and Weisheng Lu[3]

This is the authors' version of the paper:
Xue F., Guo H., and Lu W. (2020). Digital twinning construction objects: Lessons learned from pose estimation methods. In *Proceedings of the 37th Information Technology for Construction Conference* (CIB W78), São Paulo, Brazil, pp. 327–337. DOI: 10.46421/2706-6568.37.2020.paper023
The final version is freely available at: https://itc.scix.net/paper/w78-2020-paper-023

Abstract: Productivity and safety in the construction industry have long been hindered by the many uncertainties and lack of awareness in the semi-controlled site environment. The digital twinning of construction objects aims at offering digital replicas with real-time, trustable evidence for automated monitoring, human-centric decision-making, or fully automatic cyber-physical systems. This paper revisits the pose estimation methods for the digital twinning of various on-site construction objects, including construction components, equipment, and humans. From a machine learning perspective, all the pose estimation methods can be categorized into four classes, i.e., filtering, supervised, reinforcement, and unsupervised. The inputs, processes, output, and target objects of each class are introduced with demonstrative cases. Comparisons on the pros and the cons of the methods reveal the best choices for digital twinning under different objectives, such as a safer site and more productive construction, as well as constraints such as pose accuracy, computational time, and overall cost. The complexities of digital twinning different construction objects are compared to explain the distribution of existing cases in the literature. Opportunities and possible research directions in the new era of AI and blockchain are recommended at the end.

Keywords: Digital twin, Pose estimation, Machine learning, Digital construction site, Smart construction object.

1 Introduction

The construction industry has been encountering difficulties in its practices, such as endangered productivity and safety in semi-controlled site environments. Compared to other industries with controlled environments, such as manufacturing, the construction industry has seemed "backward" in the past decades (Woudhuysen and Abley 2003). Information and communication technology (ICT), such as the Internet of things (IoT), laser scanning, sensor network, and geographic information system (GIS), has been successfully applied to the monitoring and automation of construction objects, including construction components, equipment, and humans (Ahuja et al. 2009).

A digital twin is "a virtual representation of a physical object or system across its lifecycle, using real-time data to enable understanding, learning, and reasoning," according to the UK

[1] Assistant Professor, The University of Hong Kong, Pokfulam, Hong Kong SAR, China, xuef@hku.hk
[2] Associate Professor, Tsinghua University, Beijing, China, hlguo@tsinghua.edu.cn
[3] Professor, The University of Hong Kong, Pokfulam, Hong Kong SAR, China, wilsonlu@hku.hk

National Infrastructure Commission (2017). Digital twinning of construction objects, as shown in Figure 1, offers digital replicas with real-time information, which can improve the traceability and controllability of the construction objects. Digital twinning of construction objects will be promising and impactful, according to its records for other sophisticated systems, such as aircraft, wind turbines, and smart trains (Tuegel et al. 2011). In this sense, digital twinning is the equivalent to the "physical-to-cyber" subsystem of a cyber-physical system (CPS) that aims to "monitor and control the physical processes" (Lee 2008).

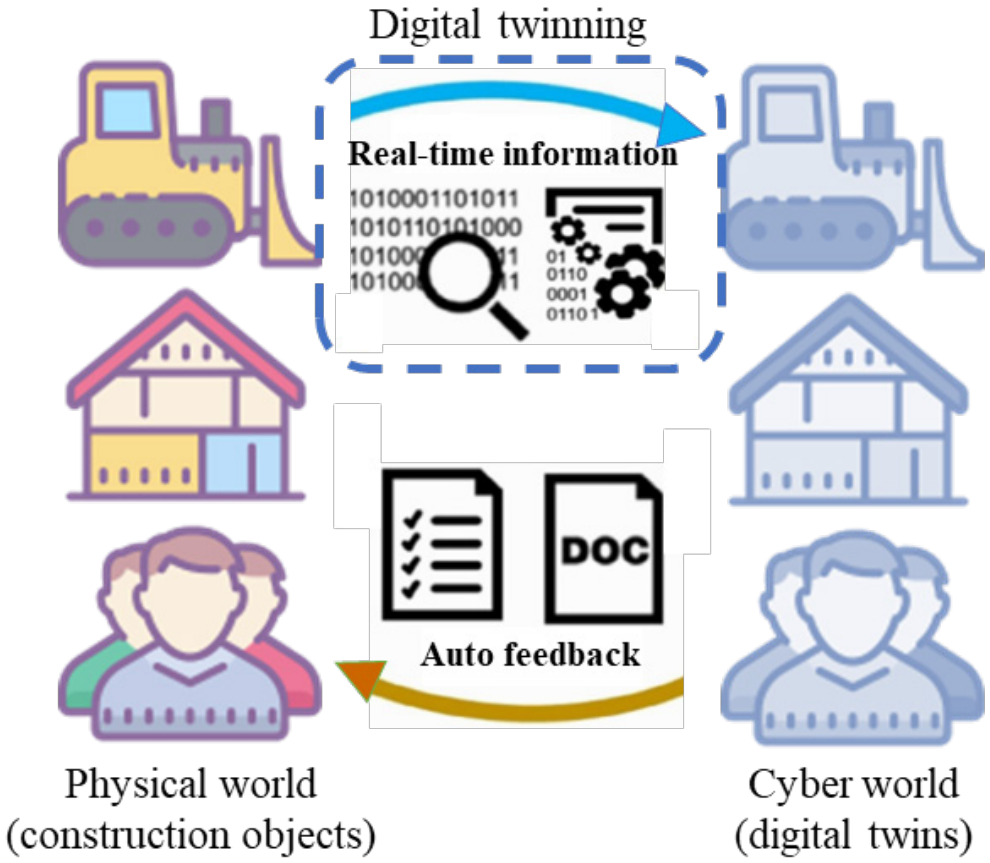

Figure 1: Digital twinning of construction objects and their positions in a cyber-physical system.

The 3D pose, as a synthesis of position, orientation, and potential purpose of a construction object, is a piece of key information for digital twinning. Pose estimation refers to the identification of such accurate position and orientation, as well as an understanding of potential purposes, from ICT sensor data for construction objects. Pose estimation methods have been investigated sporadically in construction scenarios, such as smart construction object (SCO), as-built building information model (BIM) reconstruction, construction virtual reality (VR), 4D city information model, and high-definition 3D map (Niu et al. 2016; Schwarz 2010). However, the varieties of construction objects, data inputs, and application scenarios make a one-size-fits-all method impractical. Furthermore, digital twin applications demand higher performance, for example, in near-time responsiveness and pose accuracy.

This paper aims to revisit the existing pose estimation methods in the new context of digital twinning of construction objects. In specific, a four-class taxonomy—filtering, supervised, reinforcement, and unsupervised—is borrowed from the domains of machine learning and signal processing. Section 2 briefs the research methodology. The representative methods in the four classes were demonstrated with empirical cases in Section 3. The discussion and recommendations appear in Section 4, and the conclusion is given at the end of this paper.

2 Research Methods

This study employs a comparative study on the pose estimation of construction objects. First, four models are derived from the conceptual digital twinning model in Figure 1 based on the taxonomy of machine learning methods. Then, demonstrative cases are surveyed from the literature to cover the combinations of the four classes and the three types of construction objects, i.e., components, equipment, and humans. The pros and cons of the pose estimation methods are summarized, and recommendations are given based on the demands of digital twin applications.

This study proposes a four-class taxonomy for pose estimation methods in construction, as shown in Figure 2. Due to the 'mapping-A-to-B' nature of the identification of pose information from ICT data, three classes (supervised, reinforcement, and unsupervised) are borrowed from machine learning theories. Those methods without any learning characteristics are classified as 'filtering' methods. Meanwhile, each class of method requires unique domain-specific knowledge.

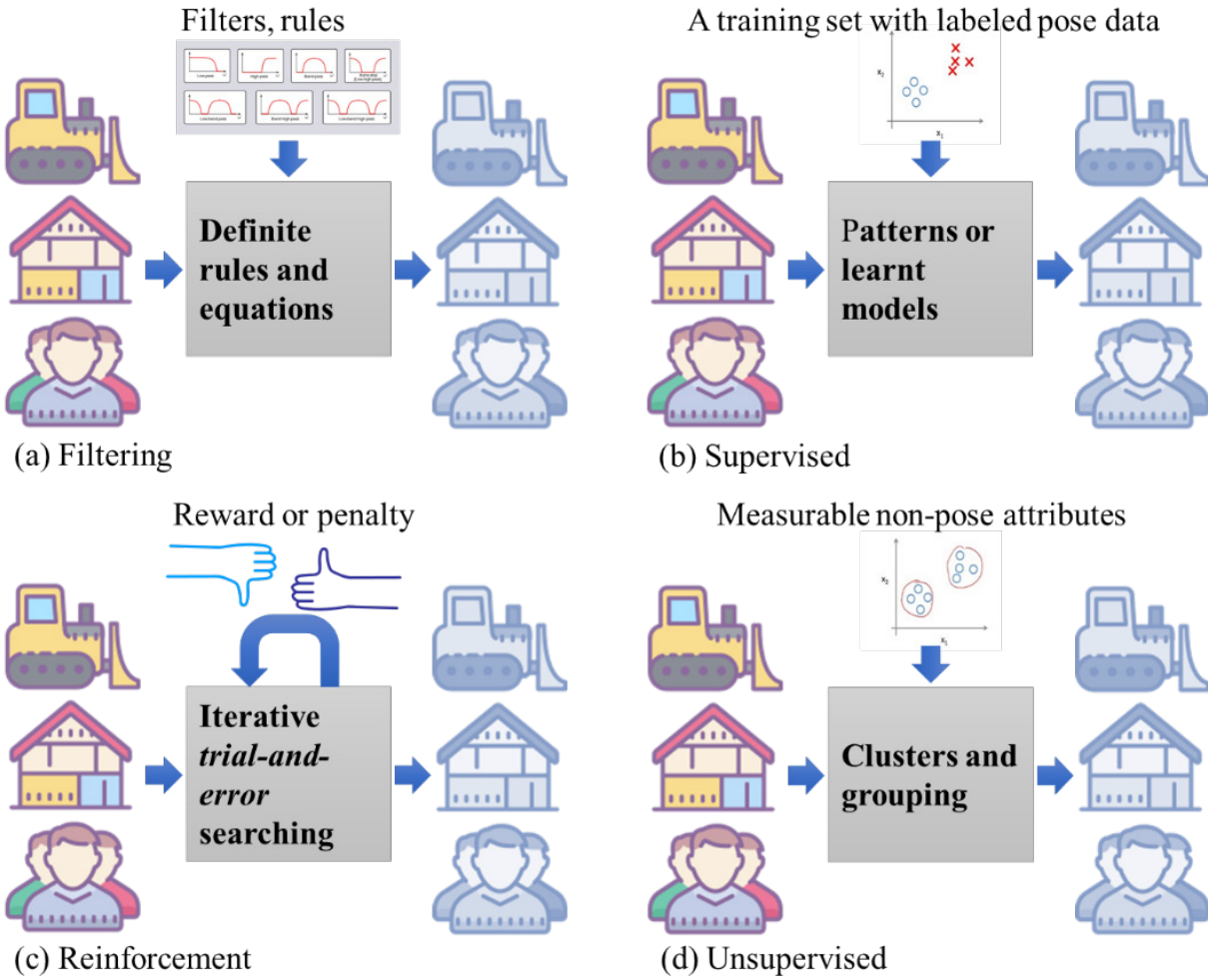

Figure 2: The four proposed classes of pose estimation methods in construction.

A literature search was conducted on Google Scholar, with the query term '(pose estimation) (construction) (human OR equipment OR facility)' and a publication date in or after 2011. The top 500 results were initially screened for the most representative cases according to the titles and abstracts within the two focused domains of construction management and computer-aided technologies (CAx). Then, the publications were furthered filtered according to the richness of information they contained on poses (e.g., position, orientation, and shreds of evidence of purposes) and ordered descending according to the average number of citations per year. A

snowballing process was then executed for the top candidate papers to include the missing query keywords. The process produced nine representative cases, as listed in Table 1, of existing pose estimation in construction.

Table 1: Representative cases of pose estimation of various construction objects.

	Components	**Equipment**	**Human**
Filtering	Niu et al. (2019)	Zhang et al. (2012)	Yan et al. (2017)
Supervised	Jin and Lee (2019)	Golparvar-Fard et al. (2013)	Han and Lee (2013)
Reinforcement	Xue et al. (2019)	–	–
Unsupervised	Kashani and Graettinger (2015)	Chen et al. (2017)	–

The cases in Table 1 covered nine out of the 12 combinations of the four classes of methods and the three types of construction objects. The missing three entries were related to the reinforcement methods for estimating equipment and human poses and unsupervised methods for human poses. Three amidst the nine papers, i.e., Zhang et al. (2012), Golparvar-Fard et al. (2013), and Han and Lee (2013), were published between 2012 and 2013, which indicated that quite a portion of pose estimation studies was established in construction before the advent of the concepts of digital twin and cyber-physical systems. Besides, many supervised learning methods that apply to construction were excluded from this study if they oversimplified the target pose information such as a 'yes/no' estimation of whether a 2D image is a worker or a truck.

3 Pose estimation of construction objects

3.1 Filtering methods

A filtering method applies fixed rules and processes, often in the form of definite rules and equations, as shown in Figure 2a. The input data contains the pose data, yet with noise and uncertainty. Based on the filter patterns or rules, the processing can result in more accurate and stable pose data. Zhang et al. (2012) developed an early real-time positioning system in 2012 based on the ultra-wideband (UWB) technology. The method, as shown in Figure 3a, employed eight sensors to measure a mobile crane's pose changes in distances of a few meters. The results showed that errors up to two meters were corrected by a 3D velocity filtering method from the raw sensor data.

Yan et al. (2017) integrated two inertial measurement unit (IMU) sensors in for estimating a construction worker's poses of head, neck, and trunk, as shown in Figure 3b. They cross-referenced the sensor data via a human backbone model and filtered and warned the 'Not Recommended' poses in real-time. For example, once the angle of trunk inclination is over 60° during manual operations, there would be a high risk of lower back pains.

Niu et al. (2019) integrated more types of sensors, including IMU, altimeter, and global positioning system (GPS) on an IoT device, for monitoring precast beam hoisting. They also applied a 3D velocity filter and identified the beam's poses and motions, including swings and rotations in the air, in real-time. The estimated poses were utilized for analyzing the near-miss safety issues as well as productivity. The 4D pose traces of the hoisted beams were visualized online in near real-time on an open GIS engine called Cesium.

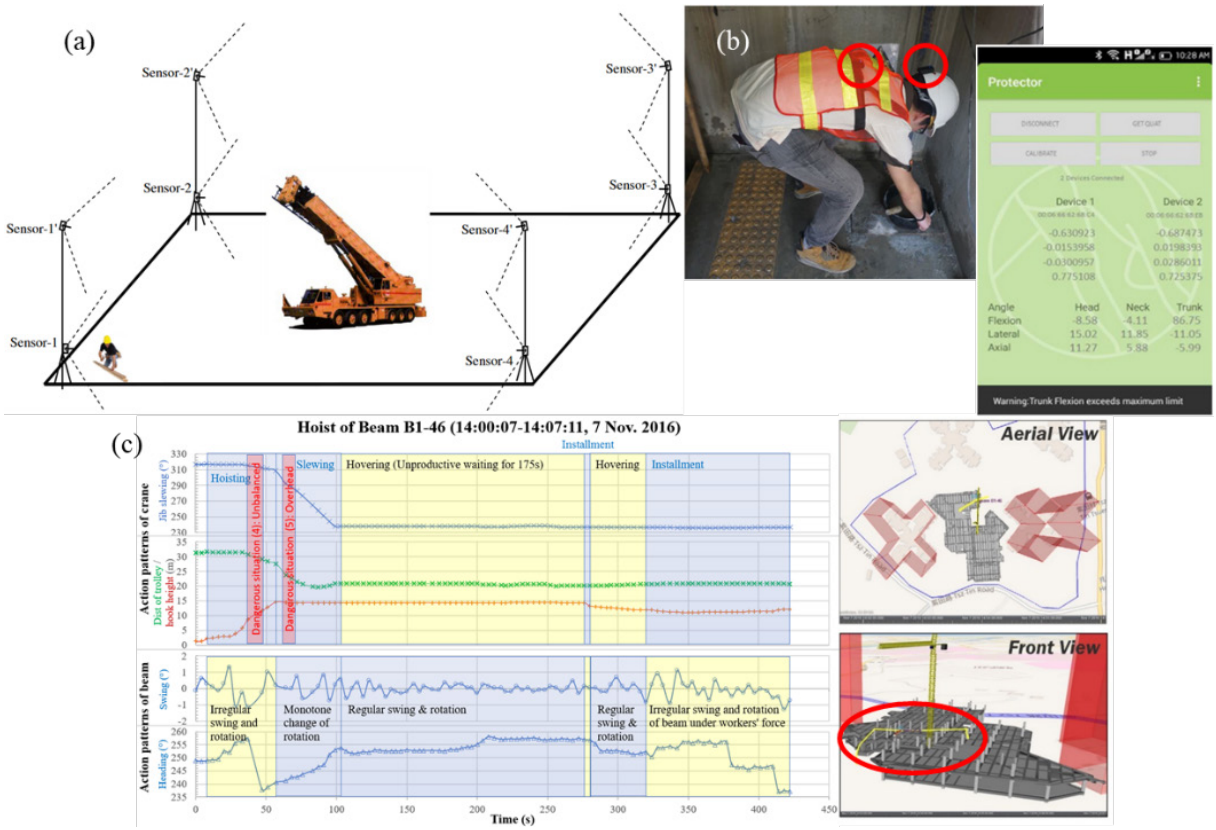

Figure 3: Filtering methods of pose estimation. (a) Sensor settings for a crane in Zhang et al. (2012); (b) Sensors (as circled) and real-time pose warnings (Yan et al. 2017); (c) Multi-sensor fusion and pose trace (as circled) in hoisting (Niu et al. 2019).

3.2 Supervised methods

A supervised method first summarizes the pose patterns or learns a meta-model from the given training data, then applies the learned patterns or models to the input data, as shown in Figure 2b. The input data table does not contain the pose data, but a list of relevant data columns called 'attributes.' Meanwhile, the training data is a table comprised of all the attributes and annotated label columns about the target pose. That is why the learned patterns or models are applicable to the input data for pose estimation. Golparvar-Fard et al. (2013) investigated the pixels' gradient-based movements and orientations in a crane video, and applied a multiple binary SVM classifiers to estimate the poses and activities, as shown in Figure 4a. They reported average accuracies at 86.33% and 98.33% for categorical actions of excavators and trucks, respectively.

Han and Lee (2013) was another early vision-based supervised method using dual cameras. The two cameras were set up to cover a target area from different view angles. First, the cameras estimated 2D poses independently through a pre-trained supervised learning model based on the Histogram of Oriented Gradient (HOG) descriptor. Then, the two 2D poses were matched into a 3D pose, as shown in Figure 4c. The recall of unsafe action detection was 88%, and the precision was also 88%.

Figure 4b shows the 3D reconstruction process of a pipeline system from a laser-scanned point cloud in Jin and Lee (2019). They applied the random sample consensus (RANSAC) to estimate the cylinder axes and employed principal component analysis (PCA) eigenvalues for distinguishing linear and curved regions based on the Catmull–Rom spline. The recall rates of the pipes were around 85% to 90%, while the computational time was about 30s for 60-90 pipes (i.e., less than 0.5s per pipe on average).

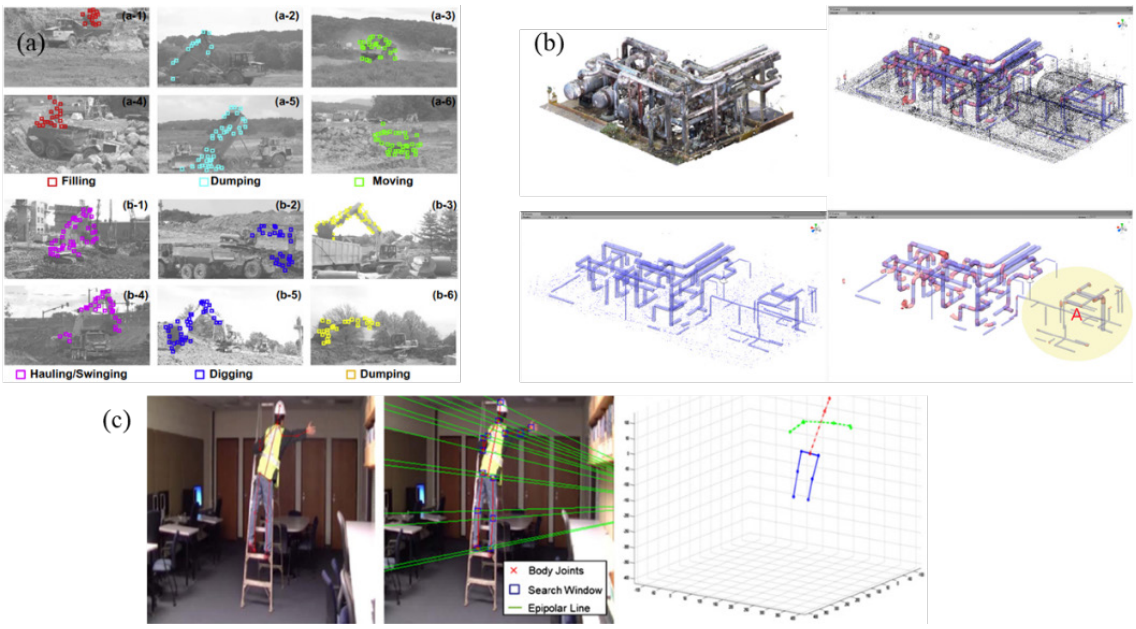

Figure 4: Supervised methods of pose estimation. (a) Learned motion features of equipment (Golparvar-Fard et al. 2013); (b) Pipelines 3D reconstruction (Jin and Lee 2019); (c) Dual-camera worker pose estimation process (Han and Lee 2013).

3.3 Reinforcement methods

In contrast, reinforcement methods were mentioned the least often in the literature. Reinforcement methods in machine learning aim to understand and automate goal-directed learning and decision-making by learning from interaction with an environment (Sutton 2018). In pose estimation, it involves repeated iterations of trial-and-error searching for the best pose, as shown in Figure 2c. Thus, a reward or penalty function is required, rather than the filters and annotated training pose data, is necessary for reinforcement methods. Xue et al. (2019) presented a reinforcement method named semantic registration based on an error function and explicit optimization algorithms, as shown in Figure 5. The test results on a 293 auditorium chairs case showed the recall and precision were around 85%, at 5.3s average computational time for each chair.

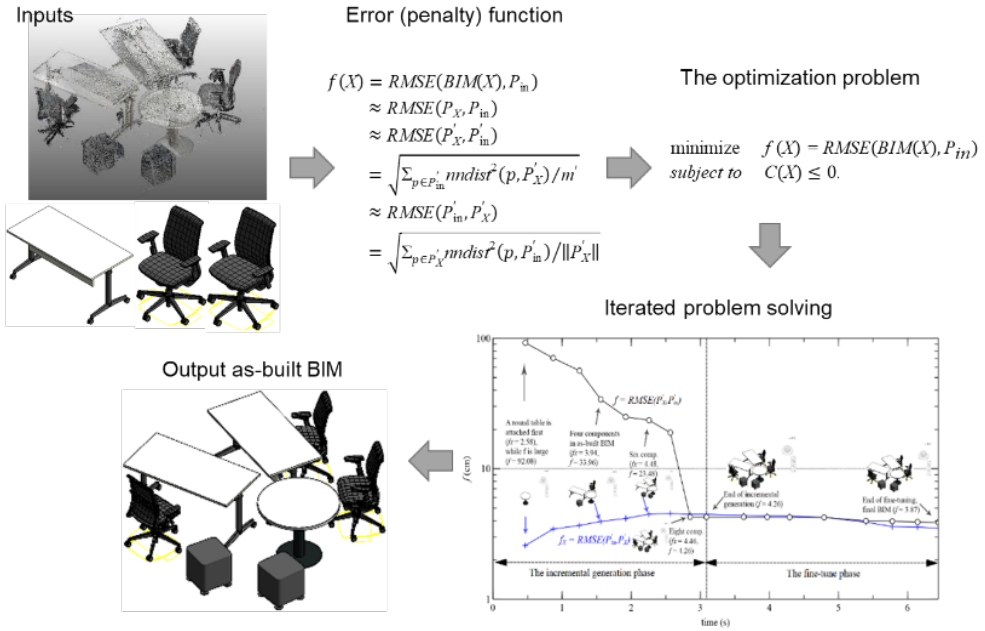

Figure 5: A reinforcement method of pose estimation in Xue et al. (2019).

3.4 Unsupervised methods

The unsupervised methods, or clustering, require the least domain knowledge in the four classes. Instead, once some non-pose attributes in the input data table are measurable, e.g., float or integer numbers, the input data records can be clustered to a few sets or a closeness-based hierarchy, as shown in Figure 2d. Although the output clusters are often not able to predict the poses directly, they are effective descriptors that recap and conceptualize a large volume of data. Thus, unsupervised methods are popular in the automated pre-processing of 3D point clouds (Chen et al. 2017; Jin and Lee 2019). Kashani and Graettinger (2015) present a direct use of the clusters for detecting rooftop damages from ground-based LiDAR data, as shown in Figure 6a. They tested the combinations of unsupervised methods and evaluation criteria and found that the Elbow method and the Calinski-Harabasz criterion resulted in an 82% correct estimation. The resulting clusters also reflected the poses of the roof elements.

Chen et al. (2017) developed a principal axes descriptor (PAD) of the cluster of points for the recognition of construction equipment. Figure 6b shows the unsupervised part. The input points were pre-processed for background removal using ground erosion. Then, Chen et al. applied an unsupervised Euclidean clustering method, where the nearest points were grouped, and isolated noise points were removed. The experimental results showed the new PAD was robust against various equipment poses. Next, after a round of supervised recognition, the precision and recall of recognizing excavators achieved over 90%, but the recall for backhoe and front loaders was no more than 75%, while the precision for bulldozers and dump trucks was no more than 60%.

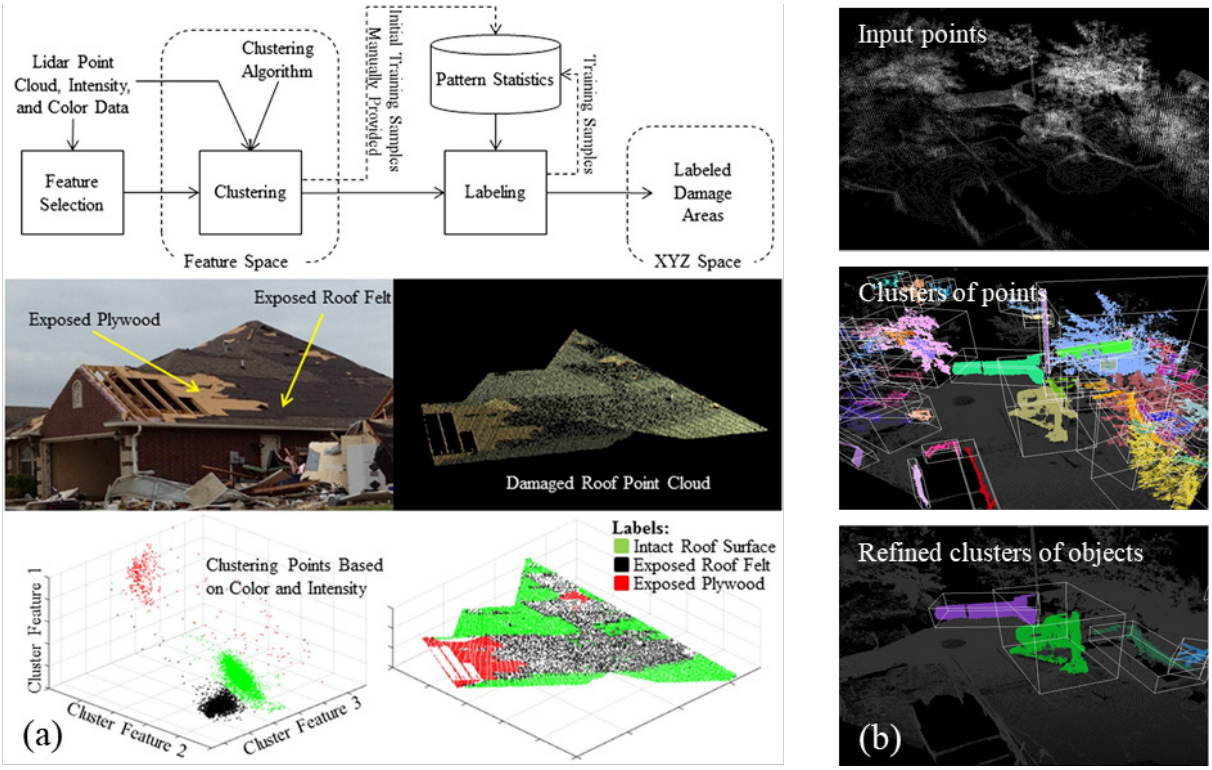

Figure 6: Unsupervised methods of pose estimation. (a) Point clusters of rooftop damages (Kashani and Graettinger 2015); (b) 3D point clusters of construction equipment (Chen et al. 2017).

4 Discussion

Pose estimation of construction objects, as well as the digital twinning, can reflect the accurate position, orientation, and potential purposes or uses. With such information being updated in a real-time fashion, the construction site environment becomes more controllable. This is, in essence, the enabler of a smart and digital construction site. Nevertheless, each class of methods discussed above has strengths and drawbacks. Table 2 summarizes the four classes of methods in terms of pose accuracy (aggregated from precision and recall), processing time, and overall cost (including hardware and staffing cost). An ideal pose estimation method should work out a high accuracy in a short time and at a low cost. However, the reality is that there is no such method.

Table 2: Comparisons of pose estimation methods for digital twinning of construction objects.

Class	Pose accuracy	Processing time	Overall cost	Example
Filtering	High	Very fast	High	Crane hoisting
Supervised	Medium	Fast	Medium	Safety supervision
Reinforcement	Medium	Slow	Low	3D reconstruction
Unsupervised	Low	Fast	Low	Data pre-processing

Table 2 shows that the filtering methods are the best in terms of accuracy and computational time, but they cost a fortune on hardware and communications. For example, an IMU-enabled IoT device costs more than US$50, and a UWB system is even higher. The reinforcement and unsupervised methods are the cheapest. However, reinforcement learning requires an in-house developed exact guiding reword function, while the unsupervised clustering results still need further processing. The supervised methods seem a balanced option. However, it requires manual labeling of the poses in the training data—which incurs additional cost. The accuracy and processing time are not acceptable in many application scenarios.

Therefore, one has to select the pose estimation methods based on their application and one's budget. It is also true for other digital twin applications in construction. If the target object is critical to the construction productivity and project delivery, for example, a tower crane or volumetric prefabricated room, then advanced sensors and filtering methods are recommended. Besides, a supervised artificial intelligence (AI) method's success should be primarily attributed not to the 'intelligence' part but to the 'artificial' part, which incurs a cost.

Furthermore, there exist three void combinations of methods and objects in Table 2. One reason is that construction equipment usually has more degrees of freedom (DoFs) than building components. For example, a mobile crane has 6 DoFs: a 3D position (x, y, z) and a 3D rotation from its chassis and the center position (x, y, z) or the equivalent (pan, tilt, length) from the jib. Human objects have even more DoFs; for instance, the human skeletal motion model in Guo et al. (2018) has 10 DoFs from the angles of the limbs, regardless of height and length. As a result, the estimation of equipment and human poses is considerably more complicated than that of building components. Plus, reinforcement and unsupervised methods are highly associated with, and thus confined to, the application scenarios, which leads to less available software libraries for the construction industry.

In the new era of AI and blockchain, new hardware and software technologies are flourishing. Some of them can be very helpful for pose estimation and the digital twinning of construction objects. For example, Birdal et al. (2018) proposed a new uniform model for

detecting all quadrics, including planes, spheres, cylinders, cones, ellipsoids, and more. Novel deep learning methods, such as Luo et al. (2019), were also efficient for pose estimation and other applications such as productivity estimation with implicit poses involved. Besides, the well-known application scenarios can also be expanded by the latest means. For example, Xue et al. (2019) applied an unsupervised method to generate a similar hierarchy of point-driven urban objects, after clustering the objects' points from aerial LiDAR data like Kashani and Graettinger (2015) and Chen et al. (2017). Penzes (2018) projected multiple application scenarios of blockchain in construction, which introduced novel distributed paradigms, real-time data exchange, transparency, and trust for general applications. It can be another direction to integrate pose estimation and digital twinning for the prospect of a smart and digital construction site.

5 Conclusion

Digital twinning of construction objects is a promising research field because it is the informational foundation for smart and digital construction site and cyber-physical construction systems that can mitigate much uncertainty and ignorance in terms of construction safety and productivity. However, it is not clear to what extent many methods fit digital twinning's purposes, such as near-time responsiveness and accuracy. This paper focuses on the pose estimation task of digital twinning and compares the pros and cons of each class of methods.

Of the proposed four-class taxonomy in this study, filtering and supervised methods are most frequently seen in the literature. The filtering methods have the best quality but also the highest cost. Although supervised methods have been significantly leveraged by recent endeavors in deep learning and big data, they can handle certain types of application scenarios. Meanwhile, reinforcement and unsupervised methods are among the cheapest, but they are complicated, closely associated with application scenarios, and sometimes require deep domain insights. Thus, the reinforcement and unsupervised methods are rarely applied to complicated (more DoFs) objects in literature. One recommended direction is to adopt the latest methods from mathematics and computer science. The other possible direction relates to the distributed paradigms, real-time data exchange, transparency, and trust that are enabled by blockchain technology.

6 Acknowledgments

The authors wish to acknowledge the collaborative support by the University of Hong Kong (No. 20300083) and Tsinghua University Initiative Scientific Research Program (No. 2019Z02HKU) and partial support by the Department of Science and Technology of Guangdong, China (Grant No. 2019B010151001).

7 References

Ahuja, V., Yang, J., Shankar, R. (2009). Study of ICT adoption for building project management in the Indian construction industry. *Automation in construction,* 18(4), 415–423. https://doi.org/10.1016/j.autcon.2008.10.009.

Birdal, T., Busam, B., Navab, N., Ilic, S., Sturm, P. (2018). A minimalist approach to type-agnostic detection of quadrics in point clouds. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3530–3540, IEEE. http://openaccess.thecvf.com/content_cvpr_2018/html/Birdal_A_Minimalist_Approach_CVPR_2018_paper.html.

Chen, J., Fang, Y., Cho, Y. K., Kim, C. (2017). Principal axes descriptor for automated construction-equipment classification from point clouds. *Journal of Computing in Civil Engineering*, 31(2), 04016058. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000628.

Golparvar-Fard, M., Heydarian, A., Niebles, J. C. (2013). Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Advanced Engineering Informatics*, 27(4), 652–663. https://doi.org/10.1016/j.aei.2013.09.001.

Guo, H., Yu, Y., Ding, Q., Skitmore, M. (2018). Image-and-skeleton-based parameterized approach to real-time identification of construction workers' unsafe behaviors. *Journal of Construction Engineering and Management*, 144(6), 04018042. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001497.

Han, S., Lee, S. (2013). A vision-based motion capture and recognition framework for behavior-based safety management. *Automation in Construction*, 35, 131–141. https://doi.org/10.1016/j.autcon.2013.05.001.

Kashani, A. G., Graettinger, A. J. (2015). Cluster-based roof covering damage detection in ground-based LiDAR data. *Automation in Construction*, 58, 19–27. https://doi.org/10.1016/j.autcon.2015.07.007.

Lee, E. A. (2008). Cyber physical systems: Design challenges. In *Proceedings of the 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC 2008)*, pp. 363–369. IEEE. https://doi.org/10.1109/ISORC.2008.25.

Luo, X., Li, H., Yang, X., Yu, Y., Cao, D. (2019). Capturing and understanding workers' activities in far-field surveillance videos with deep action recognition and Bayesian nonparametric learning. *Computer-Aided Civil and Infrastructure Engineering*, 34(4), 333–351. https://doi.org/10.1111/mice.12419.

NIC. (2017). *Data for the Public Good*. National Infrastructure Commission, UK, London. https://www.nic.org.uk/publications/data-public-good/, last accessed 2020/01/31.

Niu, Y., Lu, W., Chen, K., Huang, G. G., Anumba, C. (2016). Smart construction objects. *Journal of Computing in Civil Engineering*, 30(4), 04015070. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000550.

Niu, Y., Lu, W., Xue, F., Liu, D., Chen, K., Fang, D., Anumba, C. (2019). Towards the "third wave": An SCO-enabled occupational health and safety management system for construction. *Safety Science*, 111, 213–223. https://doi.org/10.1016/j.ssci.2018.07.013.

Jin, Y. H., Lee, W. H. (2019). Fast cylinder shape matching using random sample consensus in large scale point cloud. *Applied Sciences*, 9(5), 974. https://doi.org/10.3390/app9050974.

Penzes, B. (2018). *Blockchain technology in the construction industry: Digital transformation for high productivity*. Institute of Civil Engineers, London, UK, https://www.ice.org.uk/ICEDevelopmentWebPortal/media/Documents/News/Blog/Blockchain-technology-in-Construction-2018-12-17.pdf, last accessed 2020/01/31.

Schwarz, B. (2010). Mapping the world in 3D. *Nature Photonics*, 4(7), 429–430. https://doi.org/10.1038/nphoton.2010.148.

Sutton, R. S., Barto, A. G. (2018). *Reinforcement learning: An introduction*. 2nd edn. MIT press.

Tuegel, E. J., Ingraffea, A. R., Eason, T. G., Spottswood, S. M. (2011). Reengineering aircraft structural life prediction using a digital twin. *International Journal of Aerospace Engineering,* 2011, 154798. https://doi.org/10.1155/2011/154798.

Woudhuysen, J., Abley, I. (2003). *Why is Construction so Backward?* John Wiley & Sons, London, UK.

Xue, F., Chen, K., Lu, W. (2019). Understanding unstructured 3D point clouds for creating digital twin city: An unsupervised hierarchical clustering approach. In *Proceedings of the*

CIB World Building Congress 2019, CIB, http://frankxue.com/pdf/xue19unsupervised-preprint.pdf, last accessed 2020/01/31.

Xue, F., Lu, W., Chen, K., Zetkulic, A. (2019). From semantic segmentation to semantic registration: Derivative-Free Optimization–based approach for automatic generation of semantically rich as-built Building Information Models from 3D point clouds. *Journal of Computing in Civil Engineering*, 33(4), 04019024. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000839.

Yan, X., Li, H., Li, A. R., Zhang, H. (2017). Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention. *Automation in Construction*, 74, 2–11. https://doi.org/10.1016/j.autcon.2016.11.007.

Zhang, C., Hammad, A., Rodriguez, S. (2012). Crane pose estimation using UWB real-time location system. *Journal of Computing in Civil Engineering*, 26(5), 625–637. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000172.