Context-Aware Deep Learning Model for 3D Human Motion Prediction in Human-Robot Collaborative Construction

Xiaoyun Liang[1]; Lin Sheng[2];
Jiannan Cai, Ph.D.[3*], Shuai Li, Ph.D.[4], Yangming Shi, Ph.D.

[1]School of Civil & Environmental Engineering, and Construction Management, The University of Texas at San Antonio. Email: xiaoyun.liang@utsa.edu
[2]Department of Electrical and Computer Engineering, The University of Texas at San Antonio. Email: jeff.sheng@my.utsa.edu
[3]School of Civil & Environmental Engineering, and Construction Management, The University of Texas at San Antonio (corresponding author). Email: jiannan.cai@utsa.edu
[4]Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville. Email: sli48@utk.edu
[5]Department of Civil, Construction, and Environmental Engineering, University of Alabama. Email: shiyangming@ua.edu

ABSTRACT

Advances in robotics enable the implementation of collaborative robots in hazardous, repetitive, and demanding construction tasks to improve safety and productivity. Accurate and reliable human motion prediction is required to achieve smooth human-robot collaboration (HRC). However, many deep-learning-based models only consider observed movement to predict human motion while neglecting the interactions between humans and their surroundings. This study proposes a context-aware deep learning model, integrating observed movement and context information (i.e., locations of assigned tasks) into a Long Short-Term Memory network with an encoder-decoder architecture to predict a sequence of human motion in 3D. A pilot experiment was conducted, and the proposed model achieves an average displacement error of 0.15m. The results show that incorporating task contextual information improves the accuracy of human motion prediction by 6.25%, which could augment the perception and reasoning capability of collaborative robots for improved HRC in construction.

INTRODUCTION

Robotics has been shown to be a promising solution for safety and productivity issues existing in construction over decades (Bock, 2015). Applications of task-oriented robots have been explored in construction scenarios for the past decade, from brick-laying robots (Dörfler et al., 2016) and painting robots (Megalingam et al., 2020) to general-purpose robotic platforms for dynamic construction site conditions (Kim et al., 2021). Specifically, for human-robot collaboration (HRC) in various construction tasks (e.g. assembly) where robots work alongside human workers sharing the same workspace, collaborative robots can assist human workers in physical-demanding tasks (e.g., heavy materials delivery). To achieve safe and efficient HRC in construction tasks, it is critical for collaborative robots to understand and predict human motion accurately.

The existing studies have developed many methods for human motion prediction (Xia et al., 2022). Recurrent neural networks (RNNs) were used to predict human poses based on various activities such as walking, smoking, etc. (Martinez et al., 2017). Alahi et al. (2016) developed a

social-LSTM method for human trajectory prediction. Convolutional neural network (CNN)-based model was proposed for human motion prediction on certain movements (e.g., walking dog and greeting) by Cui et al. (2020), focusing on spatial-temporal relationships in sequences for future pose prediction. The models that are developed based on predefined actions and well-controlled scenes are not applicable to human motion prediction in construction HRC, considering that collaborative robots are exposed to unconstructed and changing workspaces. Robots need to react to changes according to their human partner's movement intention due to changing sequences and requirements of construction tasks and dynamic surroundings (Feng et al., 2015; Liang et al., 2019).

Some recent studies incorporated contextual information into human motion prediction models in manufacturing and construction applications to facilitate effective HRC in this field. For instance, Liu et al. (2019) proposed a combined network of the CNN and the long short-term memory (LSTM) network to predict repetitive work-related motions and tools required in a manufacturing computer assembly task. Zhou et al. (2022) created a gaze-data-involved framework to predict human hand motions in terms of gaze-hand clustering groups in a pipe skid maintenance task. However, those prediction models with contextual information are mainly limited to hand movement prediction, lacking an understanding and prediction of the full human body movement.

Therefore, this study proposes a context-aware deep learning model to predict 3D human motion incorporating contextual information in construction applications. This study contributes to accelerating HRC implementation in construction in two aspects. First, by integrating both observed individual movement and construction task context (i.e., the location of the assigned task), an LSTM-based network with an encoder-decoder architecture is developed to predict 3D human motion. Second, the proposed method could enable the collaborative robot to understand and reason about the human worker's movement intention according to the contextual information for adaptive robot motion planning.

METHODOLOGY

The workflow of this study is shown in Figure 1., with four steps. 1) Experiment task was designed including the locations of assigned tasks. Each participant performed the task three times, and the movements were recorded by an RGB-D camera. 2) Time-series 2D positions of human skeletons were extracted from the color images through a deep learning model. 3) To form 3D skeleton database, 3D coordinates of human skeletons as well as task locations were estimated via deprojection using depth images 4) A context-aware LSTM encoder-decoder framework was created to predict human body movement from the database obtained in Step 3. In addition, a model performance comparison was conducted afterward, where the baseline model was defined as the model without contextual information.

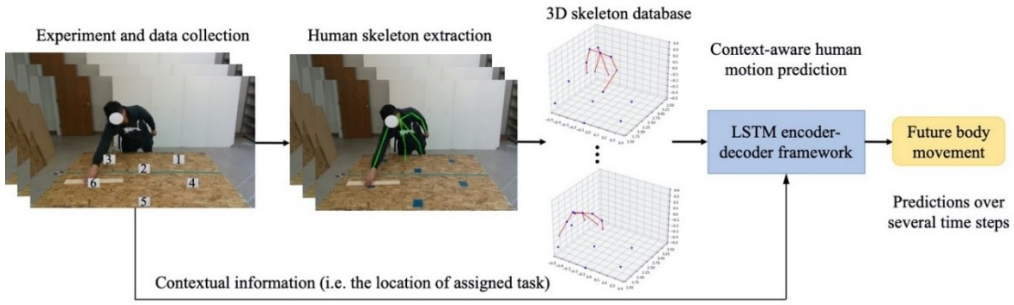


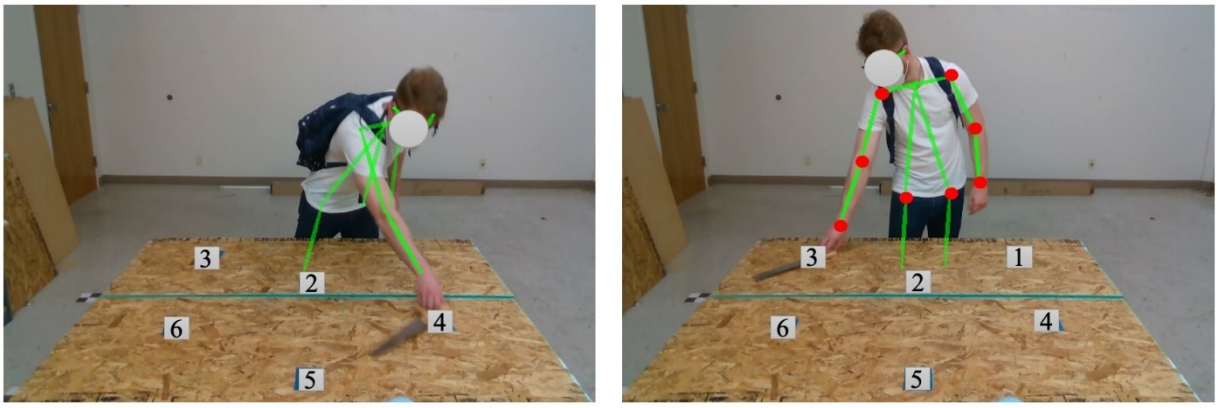
Figure 1. Overall framework

Experiment Task and Data Collection

In this study, a typical action in construction (i.e., fetch materials and tools) was designed and performed. Specifically, participants were tasked to fetch the objects (i.e., lumber and handsaw in this study) placed on a working bench. Starting with a standing position, participants would bend and reach the object one at a time and go back to the standing position. The participants performed this experiment three rounds in total, including fetching the lumber for two rounds and the handsaw for one round. In each round of experiments, the object, either lumber or handsaw, was randomly placed in one marked location on a working bench, which was repeated six times (i.e., once for each location).

The participants were only given a goal-specified description of the task, where participants could take any natural actions to complete the task. They were also asked to perform the tasks multiple times with different objects (i.e., lumber and handsaw). For example, for the task with lumber, some of the participants stood still and bent forward to reach and grab the lumber from one of the six locations. In contrast, some other participants were not able to take the same actions due to physical reasons, instead, they came closer to the side of the target location to execute the task. Such setup reflects the reality that different workers may act differently when conducting the same task in construction, which motivates the use of deep learning model to improve the generalization ability of motion prediction.

Eight participants were recruited to perform the experiment tasks. The experiment procedures were recorded by an Intel RealSense D435i RGB-D camera, containing the color and depth information of every pixel in a frame. The camera was placed and fixed in front of the workspace at a proper high.



(a) (b)

Figure 2. a) Visualization of the extracted body skeleton when the participant reaches the object placed at location #4. b) Visualization of the extracted body skeleton when the participant reaches the object placed at location #3. (Red dots present the joints.)

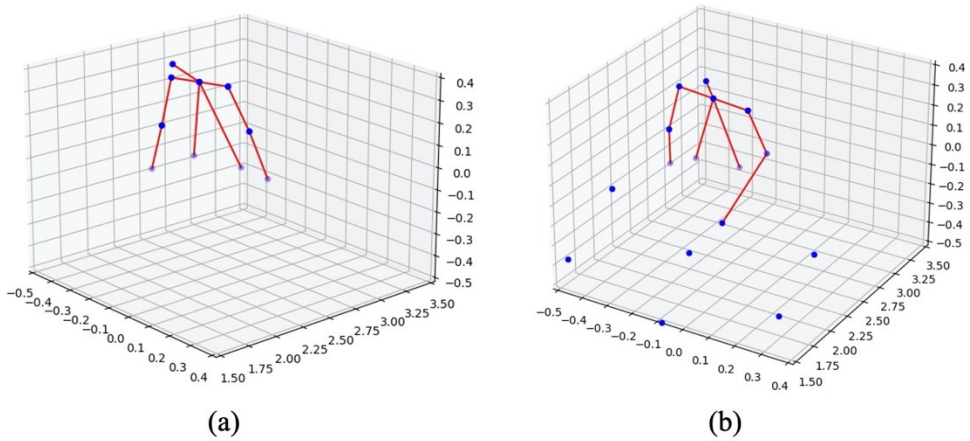
Extraction of Human Skeletons from 2D Videos

With the imagery data collected from the experiments, the human skeleton was first tracked and extracted to obtain the 2D pixel coordinates of each joint. The time-series body skeletons were extracted from the color frames, using a deep learning-based pose-tracking algorithm, Pose Flow, developed by Xiu et al., (2018). Initially, this method tracks 17 key points of the full body. Those key points are {*left ear, right ear, left eye, right eye, nose, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip*}. In our case, only the upper body was included in the skeleton database because the table blocked the lower body of the participant. Additionally,

the positions of the eyes, nose, and ears yield redundant information in our case, because this study focuses on the full human body skeleton mainly. As a simplification, the utilization of the mean value of left and right eye data was adopted as a representative feature for the head in this study. As shown in Figure 2, nine key points are considered in the extracted body skeleton, including {*head, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip*}. The 2D pixel location of each key point was extracted and stored in terms of time-series frames. Figure 2. shows an example that participants reach the objects at different locations on the bench.

3D Human Skeleton Database

Given the recorded data containing the color images and depth information for every pixel, the 3D coordinates of each key point of the human upper body were extracted. Utilizing the 2D pixel locations of the joints extracted in the second step, deprojection was performed using the Intel RealSense SDK, which converts a 2D pixel location in an image to a 3D point location indicating the depth of this point. Due to hardware capacity, the frame rate was set as 15 fps for video recording, which may result in missing frames of movements when the participant executed quickly. Furthermore, in the object-fetching task, occlusion of some body parts often occurs, where RGB-D camera experiences difficulty in capturing the depth of the full body. For example, Figure 2. (a) shows that the left arm of the participant was obscured by the body. Thus, to build a valid 3D human skeleton database, the 3D coordinates of human skeletons that are converted from 2D pixel locations need to be cleaned and processed.



(a) (b)

Figure 3. a) Visualization of 3D Human skeleton. b) Visualization of converted nine key points of the human skeleton and six locations in 3D coordinates.

First, the outliers along z-axis (i.e., depth direction) are removed from the dataset, whose value is greater than the maximum distance that a participant moves around from the camera. Then, the missing points are filled with interpolation, where they are gradually filled in terms of the previous and following valid data. After the data processing, the data is visualized to check its validity by considering the physical constraints of the limbs and torso. Finally, meaningless skeletal structures (e.g., the skeletal poses are not realistic nor feasible considering physical constraints) are removed from the 3D human skeleton database. Furthermore, given the assigned task of each video, the location coordinates of each task are added accordingly. Thus, the 3D human skeleton database is completed. In this study, the time-series 2D pixels of nine key points

were converted to 3D coordinates (see Figure 3. (a)). Along with the joint coordinates, the pixels of six possible task locations were also extracted and converted to 3D coordinates (see Figure 3. (b)).

Context-aware Deep Learning Model

This study proposes a context-aware LSTM-based model with encoder-decoder architecture to integrate location coordinates of assigned tasks into 3D human motion prediction (see Figure 4). The 3D coordinates of nine joints obtained from RGB-D camera are normalized and shaped into a 27-dimensional vector. It is concatenated with the locations of assigned tasks for task-related context integration. The 30-dimensional features over several time steps are fed into LSTM encoder, and the decoder predicts the following several time steps of human movement according to the dynamics of human motion that the model captured. The loss function of this model is mean squared error (MSE) loss for human motion prediction. A similar architecture is proved robust and efficient in worker trajectory prediction on construction sites (Cai et al., 2020). Specifically, the proposed method focuses on integrating task-related context and 3D human motion to predict 3D human body movement.

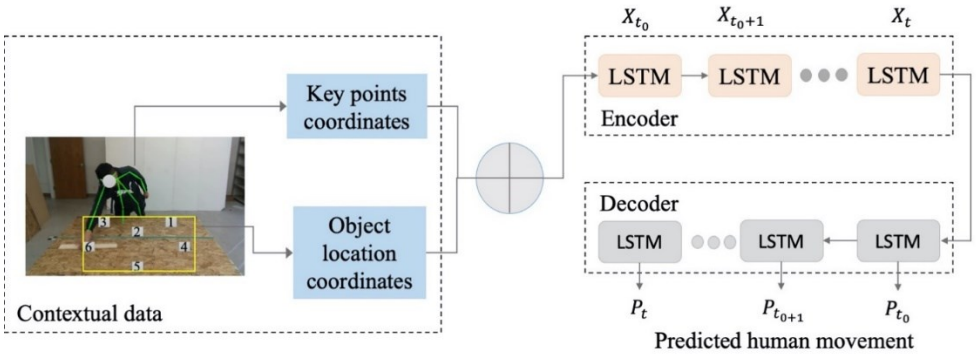


Figure 4. Context-aware prediction model

IMPLEMENTATION

The proposed framework was trained and tested on the 3D human skeleton datasets collected from the experiment. A 2D skeleton dataset was first constructed, which contains 126 movement sequences of six different tasks (i.e., fetch the object from six different locations on the table) performed by seven different participants. Each task was performed three times, including fetching lumber twice and the handsaw once. Thus, a total of 11,340 frames (at 15 fps) were collected from the RGB-D camera. Utilizing the 2D pixel locations, 3D coordinates were recovered using the deprojection function provided by Intel RealSense SDK 2.0. After data processing, results from some frames were excluded from the database due to meaningless skeletal structures, resulting in 3D point coordinates of 9,257 frames in total. One important reason for having awkward poses is that half of the upper body could be blocked when the participant is in a twist pose to execute the task (e.g. when the participant uses the right hand to fetch the object from locations 1 and 6). This makes it impossible for the RGB-D camera to capture the blocked side of the body. Some frames contained good results on the visible side of the body, and the values were kept while results of the blocked side were replaced with zero. Table 1 shows the distribution of frame number across different object locations.

Table 1. Number of frames for each task based on different object locations

Location	1	2	3	4	5	6
Number of frames	1320	1611	1507	1572	1646	1601

Following the study conducted by (Mao et al., 2019), 400ms (i.e., 6 frames) was used in both observation duration and prediction duration for short-term prediction. After applying the sliding window with a stride of 2 to the dataset, the sample size became 3,432. 80% of the dataset was used in model training, and the rest of 20% was used for testing. In the training process, a 5-fold cross-validation was adopted to tune the parameters, including the number of epochs, batch size, and hidden size. The average performance of the 5 models based on 5 sub-datasets indicated the performance of a group of parameters, which was used to select the best parameters for this study. After the selection was completed, the model was retrained using all training data and was tested on the testing dataset for model performance evaluation. The hyperparameters selected were batch size of 20, hidden size of 70, and epoch number of 1000. The optimizer used in this study was Adam, with a default learning rate of 0.001.

RESULTS

The performance of this model was evaluated using two commonly used metrics (Yuan & Kitani, 2020): average displacement error (ADE), which measures the mean of the Euclidean distance between the ground truth and predictions of joint positions over predicted time steps; and final displacement error (FDE), which measures the Euclidean distance between final predicted joint positions and the true final joint positions. These two metrics can be expressed as:

$$ADE = \frac{\sum_{i=1}^{N}\sum_{t=0}^{T}\left\|\tilde{y}_t^i - y_t^i\right\|}{N \times T} \qquad (1)$$

$$FDE = \frac{\sum_{i=1}^{N}\left\|\tilde{y}_t^i - y_t^i\right\|}{N} \qquad (2)$$

where $\tilde{y}_t^i$ and y_t^i are the predicted and the ground-truth joint positions of data i at time t, N is the number of data samples, and T is predicted time steps.

The results were compared with baseline method to assess the impact of integrating contextual information into human motion prediction. The baseline model has the same structure in terms of motion prediction, but the input features do not include location coordinates of assigned tasks. The proposed model integrates contextual information in observation period (i.e. with location coordinates of assigned tasks). Table 2 presents the comparison results. Both models result in a relatively high accuracy – with both ADE and FDE less than 0.2 m. Context-aware model has slightly improved the performance by 6.25% in ADE and 5% in FDE compared to the baseline model.

Table 2. Evaluation results of human motion prediction

Model	**ADE** (m)	**FDE** (m)
Baseline model	0.16	0.20
Context-aware model (ours)	0.15	0.19

CONCLUSION

This study proposes a context-aware deep learning model, which integrates both observed individual movement and construction task context (i.e., the location of assigned task) into an LSTM network with an encoder-decoder architecture, to predict a sequence of future 3D human body movement. The method was tested and validated using experimental data collected from seven participants via an RGB-D camera. 3D human skeletons were extracted from video recordings, and a corresponding database was formed. The model was quantitively evaluated and achieved an ADE of 0.15 m and an FDE of 0.19 m. The model was further compared with a baseline model that does not incorporate contextual information, resulting in a 6.25% improvement in ADE and a 5% improvement in FDE.

For this study, the limitations remain. First, a single type of movement is included in the dataset, which could limit the model performance on generalizability. To improve and test the robustness of this model, new persons' data and data from diverse construction activities should be included in further study. Progressively, to enhance the implementation of HRC in construction, the dataset should involve human-robot collaborative tasks and the robot's movement as well. As an ongoing effort, the dataset is continuously enriched with various HRC construction tasks and robot-related data. Second, human motion was recorded using one RGB-D camera in this study, which was subject to some uncertainties due to occasional occlusion and low frame rate. Although the data were processed to mitigate such uncertainties in this study, an advanced motion-tracking system could be used to provide more accurate and robust ground-truth data. Third, the improvement of performance when incorporating contextual information is not significant compared to the baseline model. The main reason could be the relatively small sample size, considering some human movement data from certain locations were missing in the dataset due to occlusion. The proposed method will be further evaluated using an enriched dataset, as indicated in the first aspect.

ACKNOWLEDGMENTS

This research was funded by the U.S. National Science Foundation (NSF) via Grants 2138514, 2222670, and 2222810. The authors gratefully acknowledge NSF's supports. Any opinions, findings, recommendations, and conclusions in this paper are those of the authors, and do not necessarily reflect the views of NSF, The University of Texas at San Antonio, The University of Tennessee, Knoxville, and University of Alabama.

REFERENCES

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). *Social LSTM: Human Trajectory Prediction in Crowded Spaces*. 961–971. https://openaccess.thecvf.com/content_cvpr_2016/html/Alahi_Social_LSTM_Human_CVPR_2016_paper.html

Bock, T. (2015). The future of construction automation: Technological disruption and the upcoming ubiquity of robotics. *Automation in Construction*, *59*, 113–121. https://doi.org/10.1016/j.autcon.2015.07.022

Cai, J., Zhang, Y., Yang, L., Cai, H., & Li, S. (2020). A context-augmented deep learning approach for worker trajectory prediction on unstructured and dynamic construction sites.

Advanced Engineering Informatics, *46*, 101173. https://doi.org/10.1016/j.aei.2020.101173

Cui, Q., Sun, H., & Yang, F. (2020). *Learning Dynamic Relationships for 3D Human Motion Prediction*. 6519–6527. https://openaccess.thecvf.com/content_CVPR_2020/html/Cui_Learning_Dynamic_Relationships_for_3D_Human_Motion_Prediction_CVPR_2020_paper.html

Dörfler, K., Sandy, T., Giftthaler, M., Gramazio, F., Kohler, M., & Buchli, J. (2016). Mobile Robotic Brickwork. In D. Reinhardt, R. Saunders, & J. Burry (Eds.), *Robotic Fabrication in Architecture, Art and Design 2016* (pp. 204–217). Springer International Publishing. https://doi.org/10.1007/978-3-319-26378-6_15

Feng, C., Xiao, Y., Willette, A., McGee, W., & Kamat, V. R. (2015). Vision guided autonomous robotic assembly and as-built scanning on unstructured construction sites. *Automation in Construction*, *59*, 128–138. https://doi.org/10.1016/j.autcon.2015.06.002

Kim, S., Peavy, M., Huang, P.-C., & Kim, K. (2021). Development of BIM-integrated construction robot task planning and simulation system. *Automation in Construction*, *127*, 103720. https://doi.org/10.1016/j.autcon.2021.103720

Liang, C.-J., Lundeen, K. M., McGee, W., Menassa, C. C., Lee, S., & Kamat, V. R. (2019). A vision-based marker-less pose estimation system for articulated construction robots. *Automation in Construction*, *104*, 80–94. https://doi.org/10.1016/j.autcon.2019.04.004

Liu, Z., Liu, Q., Xu, W., Liu, Z., Zhou, Z., & Chen, J. (2019). Deep Learning-based Human Motion Prediction considering Context Awareness for Human-Robot Collaboration in Manufacturing. *Procedia CIRP*, *83*, 272–278. https://doi.org/10.1016/j.procir.2019.04.080

Mao, W., Liu, M., Salzmann, M., & Li, H. (2019). *Learning Trajectory Dependencies for Human Motion Prediction*. 9489–9497. https://openaccess.thecvf.com/content_ICCV_2019/html/Mao_Learning_Trajectory_Dependencies_for_Human_Motion_Prediction_ICCV_2019_paper.html

Martinez, J., Black, M. J., & Romero, J. (2017). *On human motion prediction using recurrent neural networks* (arXiv:1705.02445). arXiv. https://doi.org/10.48550/arXiv.1705.02445

Megalingam, R. K., Prithvi Darla, V., & Kumar Nimmala, C. S. (2020). Autonomous Wall Painting Robot. *2020 International Conference for Emerging Technology (INCET)*, 1–6. https://doi.org/10.1109/INCET49848.2020.9154020

Xia, X., Zhou, T., Du, J., & Li, N. (2022). Human motion prediction for intelligent construction: A review. *Automation in Construction*, *142*, 104497. https://doi.org/10.1016/j.autcon.2022.104497

Xiu, Y., Li, J., Wang, H., Fang, Y., & Lu, C. (2018, February 3). *Pose Flow: Efficient Online Pose Tracking*. ArXiv.Org. https://arxiv.org/abs/1802.00977v2

Yuan, Y., & Kitani, K. (2020). DLow: Diversifying Latent Flows for Diverse Human Motion Prediction. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 346–364). Springer International Publishing. https://doi.org/10.1007/978-3-030-58545-7_20

Zhou, T., Wang, Y., Zhu, Q., & Du, J. (2022). Human hand motion prediction based on feature grouping and deep learning: Pipe skid maintenance example. *Automation in Construction*, *138*, 104232. https://doi.org/10.1016/j.autcon.2022.104232