Towards Part-Based Construction Equipment Pose Estimation Using Synthetic Images

Mohammad M. Soltani[1]; Zhenhua Zhu[2]; and Amin Hammad[3]

[1]Dept. of Building, Civil and Environmental Engineering, Concordia Univ., 1515 Ste-Catherine St. West, EV9.415, Montreal, QC, Canada H3G 2W1. E-mail: mo_solta@encs.concordia.ca

[2]Dept. of Building, Civil and Environmental Engineering, Concordia Univ., 1515 Ste-Catherine St. West, EV6.237, Montreal, QC, Canada H3G 2W1. E-mail: zhzhu@bcee.concordia.ca

[3]Concordia Institute for Information Systems Engineering, 1515 Ste-Catherine St. West, EV7.634, Montreal, QC, Canada H3G 2W1. E-mail: hammad@ciise.concordia.ca

Abstract

Monitoring the pose of the construction equipment is an essential prerequisite for determining safety and productivity of construction processes. In order to make construction sites safer, there is high demand for accurate and real-time motion tracking tools and methods to capture any movement of the equipment and its parts. Computer vision (CV) techniques are becoming more popular because of their lower cost of deployment and availability compared with other techniques. However, few CV methods have been focused on equipment part detection and pose estimation. This paper aims to propose a new method to detect the parts of the construction equipment that can be used to detect its pose. Using the concept of synthetic images for each equipment's part (e.g., bucket, dipper, boom, and body), multiple detectors are trained for each part from different views and applied to recognize the parts. The synthetic images are generated by overlaying the images of the 3D model of the equipment on the real images of the construction sites as background.

1 INTRODUCTION AND BACKGROUND

Monitoring the safety and productivity on the construction site is a difficult task since it requires a lot of effort for collecting and processing the related data. One of the most important sources of information for evaluating the safety and productivity of the construction equipment is the near real-time pose data of the equipment. Estimating the pose of articulated equipment (e.g. excavator) depends highly on the correct detection of the parts. Focusing on excavators, this research aims to investigate the potential of detecting the equipment parts as a fundamental step toward pose estimation by developing a comprehensive database of synthetic images of the excavator and its parts from various views with different scales and light conditions and applying the detectors trained using the mentioned database on the images and video frames recorded from the construction sites. In the following section, previous studies that focused on equipment pose estimation are reviewed.

Azar and McCabe (2012) described a novel approach for articulated equipment recognition and pose estimation. They reviewed the concept of the latent

Support Vector Machine (SVM) part-based models (Felzenszwalb et al., 2010) which applies a classifier to detect the whole body of the equipment as root and then looks for the body parts within the root. Since the shape of whole excavator body is changing significantly over time, they suggested using a major part of the body as root and searching for the adjacent parts around the root. The boom of the excavator was considered as root, and the dipper was selected as the adjacent part. For each side of the excavator (only right and left), they trained three classifiers (in total six for both sides), which cover the dipper in three positions (horizontal, inclined, and vertical). Furthermore, a spatio-temporal reasoning method (Renz & Nebel, 2007) was applied to improve the detection rate. This method used a logical relationship between the detected bounding boxes around the target object over time to reject the false detections. The results of this research show an accuracy of 95.2%; however, this method is only applicable in videos with static backgrounds.

Recently, Azar et al. (2015) proposed a new framework for monitoring the motion of the excavator's boom and dipper using planar markers attached to the boom and dipper. Figure 1 shows how the AprilTag (Olson, 2011) markers are attached to the excavator parts. To determine the axis of each part, at least two markers are required to be placed for each part. The results of their sensitivity analysis reveal that although this method provides a reasonable accuracy and is considered as a low-cost and user-friendly method, it is unable to track 3D poses and requires a clear line of sight of the parts, clean and visible markers, and the camera plane to be parallel to the marker plane.

Figure 1. Marker based excavator pose estimation (Azar et al., 2015)

An early research for understanding construction activities was done by Yang et al. (2011). The ultimate goal of this research was to find the current activity of tower cranes out of two categories: concrete pouring and non-concrete material movement. They applied the 2D-3D rigid pose estimation and the silhouette-based tracking algorithms to estimate the jib angle and the trolley position.

A video interpretation model was proposed by Gong and Caldas (2010) to assess automatically the productivity of construction operations. The method tries to synchronize the operation happening in the video with the corresponding simulation model of the same operation. The main data that their method requires from the video

are the spatial, temporal, and semantic contexts. Having the mentioned data ready on hand, their model can detect the event, classify its state, and recognize the ongoing scenario. Later on, Gong et al. (2011) extended the previous research by using Bag-of-Video-Feature-Words and Bayesian network models to learn and classify the actions of the workers and the equipment

Another research done by Golparvar-Fard et al. (2013) presented an approach for learning and classifying the equipment actions. The method extracts the spatio-temporal features of the equipment actions from the video frames of the related construction processes. These features are then used to generate the Histogram of the Oriented Gradient (HOG) feature descriptors to create the code words using k-Means clustering technique. A multi-class SVM classifier learns the distribution of the code words and applies it to the new videos to recognize the equipment action classes. The proposed method was tested on the videos containing an excavator and a truck, and the accuracies of 86% and 93% were achieved, respectively.

Furthermore, a framework named Server-Customer Interactive Tracking (SCIT) was proposed by Azar & McCabe (2013) and extended by Azar et al. (2013). The system, which consists of image and video processing module, spatio-temporal reasoning module, and action recognition module, can detect and track the machines and recognize their activities. The worst accuracy provided from eight site visits was 86%, and the best one was 100% while the average was 95%.

Soltani et al. (2015) investigated the approach of annotating the images automatically using a 3D model of the object to recognize in the images of construction sites. The results showed that the proposed automatic annotation method using synthetic images can play the role of real images captured from the construction site for training purposes. Moreover, the automatic annotation significantly reduced the required time for defining Region of Interest (ROI) more than 90% compared to traditional annotation methods while the accuracy of the object recognition is improved by training more synthetic images.

2 METHODOLOGY

Reviewing the literature motivated the authors to propose a new method for estimating the pose of the construction equipment. The method can be applied on single images whether they are static images or separated frames of a video. In other words, in the proposed method, the temporal information of the target objects in the preceding and succeeding frames is not involved in the process for estimating the pose of the equipment. The framework of the proposed method is presented in Figure 2. Training the equipment parts' detectors requires positive image dataset and negative image dataset. The positive images are created using the concept of the synthetic images proposed by Soltani et al. (2015) and the part detectors are trained using the generated positive and negative samples.

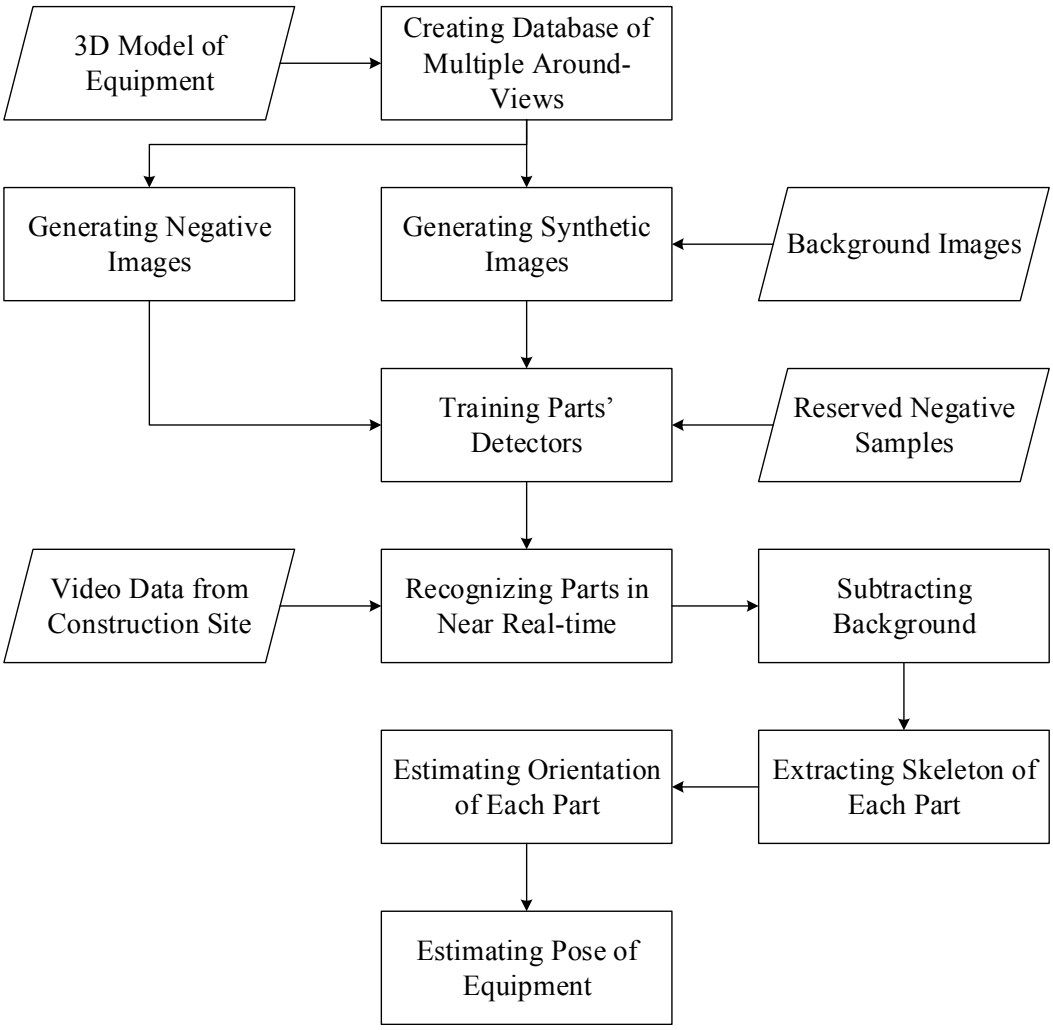

Figure 2. Process of recognizing equipment parts

The synthetic images based on the 3D model of construction equipment can be used for training the detector, which can recognize the target object within a new dataset. The trained detectors are then applied on the cropped frames to find the parts of the equipment. The following sections explain the details of the generation of positive and negative images.

2.1 Generation of Positive Images

The 3D model of the equipment is assumed to be available for creating the synthetic images. Using the method proposed by Soltani et al. (2015), the around-view images of the equipment with a single color background are created within the 3D modeling tool by the server. These images are then used for creating the synthetic images by integrating the real images of the construction sites as their background. Additionally, each part of the equipment is used for generating the part-based image datasets where the other parts are hidden. This process results in four datasets for the excavator's parts (i.e. dipper, boom, body, and bucket).

The process of generating and annotating the synthetic images for the parts would be slightly different compared to the similar process explained in Soltani et al. (2015) as shown in Figure 3. First, the image of each part with the single color background is segmented to recognize the part, and a bounding box is drawn around each part (Figure 3(a) and Figure 3(b)). Afterward, the generated boxes are added to the image of the whole equipment (Figure 3(c)) from the same view, and the background image (Figure 3(d)) is added to the image. As result, Figure 3(e) shows the annotated image of the equipment's part.

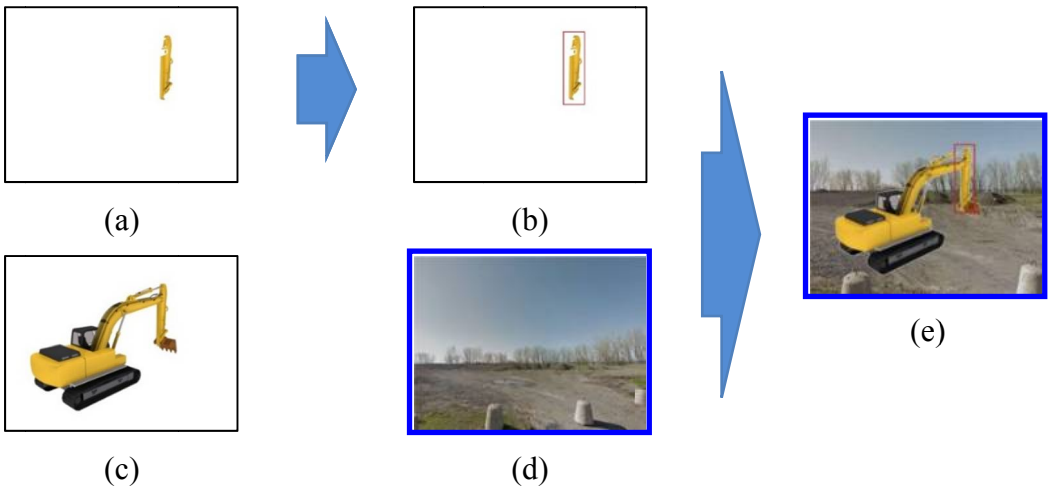

Figure 3. Parts auto annotation process

2.2 Generation of Negative Images

Another important step is to prepare the negative images for the training phase. In addition to the auto-generated negative images explained in Soltani et al. (2015), another set of images needs to be prepared to help the part detectors to differentiate each part of the equipment from other parts. Therefore, after generating the bounding boxes around the parts in the previous steps, the content of each bounding box is stored as a negative sample for the training of the other parts. For instance, the cropped areas from the synthetic images of the boom, bucket, and body of the excavator are used as negative samples for the training of the dipper. In the next phase, HOG algorithm is applied to train the detectors for each part of the equipment using the created positive and negative samples. The detectors are sent to the main server for recognizing the parts of the excavator (more information on this process is given in Soltani et al. (2015)). The bounding boxes of the detected areas for each part are extracted by applying the parts' detectors individually.

3 CASE STUDY

The proposed method is validated by performing the experiments on the static image dataset and the consecutive frames extracted from a video. The 3D model of the excavator plays a key role for generating the images of the excavator and its boom, dipper, bucket, and body. After adjusting the light and the reflection conditions of the model taken from Google Warehouse in Autodesk 3D Maxs, 17 virtual

cameras were defined as shown in Figure 4. These cameras are located every 10 degrees from the equatorial plane of the sphere.

The covering area starts at 85˚ on the north and finishes by 75˚ on the south. Each camera traverses on a horizontal circle with a step of 1˚ focusing on the excavator. The first dataset was created using the captured images from the whole excavator while the background is set to white color. In the next step, the color of all parts except the target part was changed to the same white color as the background (e.g. if the target part is the body, the color of the boom, bucket, and dipper was changed to white) and then the around-view images were captured. The reason for making the parts white in the proposed method instead of hiding them is that these parts can block the full view of the target part, and this fact needs to be considered (e.g. the boom was blocking the cabin of the operator as shown in Figure 5(b)). As shown in Figure 5(a), the boom, dipper, and bucket were hidden while in Figure 5(b) the boom, dipper, and bucket were whitened.

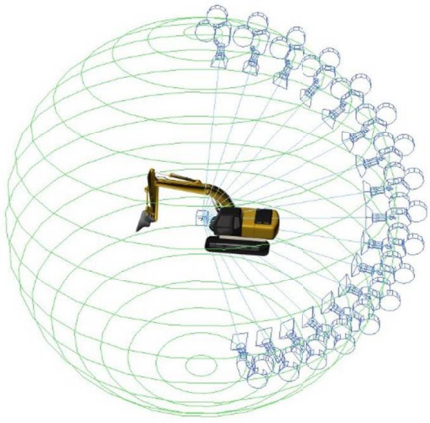

Figure 4. Spherical position of the camera

(a) Hidden boom (b) Whitened boom

Figure 5. Equipment image generation alternatives

The architecture of the database of the equipment images is illustrated in Figure 6. The database consists of five sub-databases for each part (the body, dipper, boom, bucket, and all parts together). Within each dataset, 17 datasets are created for every 10˚ latitude starting at +85˚ and finishing by -75˚.

Each dataset covers the around-view of the parent database at the specific latitude every one degree. The detailed information regarding the data stored in the excavator database is provided in Table 1. Each dataset has 360 images, and its parent sub-database includes 17 datasets with 6,120 images in total. The main database has five sub-databases and 30,600 raw images.

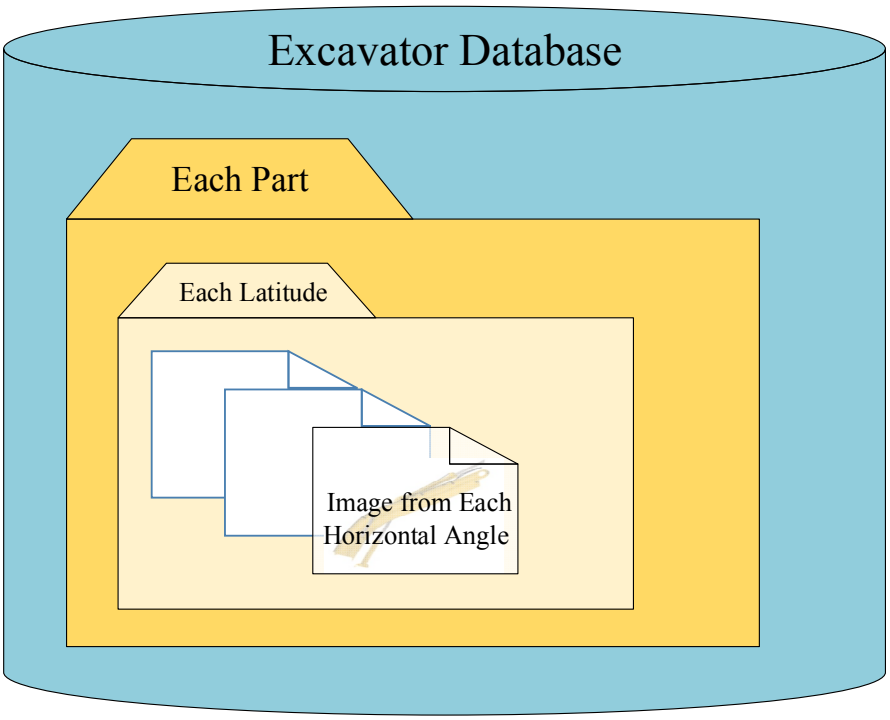

Figure 6. Equipment image database structure

Although this database covers all views of the equipment, some views may not be used for the training phase, or they may be used only in special cases. For instance, the view of the equipment at -25˚, -35˚, -45˚, -55˚, -65˚ or -75˚ can be used in a case when the equipment is working on the top of a hill, but the camera is located on the ground down to the hill. Therefore, all views are stored in the database for any further possible application.

Table 1. Equipment image database layers

Data Structure	Excavator Database	Sub-Database	Dataset
Content	5 Sub-Databases	17 Datasets	360 Images
Number of Images	30,600	6,120	360

3.1 Experiment on Static Images

In the first test, three detectors are trained and evaluated for each part, and the result of each detector is compared with the other detectors. In all trained detectors, 15 backgrounds, including one white color background, are used to generate the synthetic images. Moreover, three lighting conditions are applied to the created images. For each part, three cases are considered to be evaluated. The first detector is trained using the images viewed from the latitude of -5˚ to +55˚ and 45˚ to 135˚ on the horizontal axis considering one size of the object which is the largest possible size of the equipment in one image. The total generated synthetic images are 28,665 (91 images from one-quarter of the around-view multiplied by 3 light conditions, 15

backgrounds, and 7 latitudes). The second case covers from the latitude of +5˚ to +45˚ and 45˚ to 135˚ on the horizontal axis considering one size of the equipment with 20,475 images (91 images from one-quarter of the around-view multiplied by 3 light conditions, 15 backgrounds, and 5 latitudes). The third case contained 61,425 images from latitude of +5˚ to +45˚ and 45˚ to 135˚ on the horizontal axis considering three sizes of the object which are the largest, half, and quarter size of the equipment in one image (91 images from one quarter of the around-view multiplied by 3 light conditions, 15 backgrounds, 5 latitudes, and 3 sizes). This process is repeated three times for the dipper, boom, and body. The bucket results were for not reliable because the bucket is usually covered by dirt and soil and it is very hard to differentiate it from the soil.

Regarding the negative samples, the auto negative sampler generates 1,600 images randomly out of 14 backgrounds (15 minus one white color background) as the fixed negative dataset. Moreover, for case one, 114,660 images including 85,995 images of the other parts except the target part and 28,665 images of the positive images while the target part is deleted from them. The variable numbers of the negative images follow the same procedure as in case one and are added to the fixed negative samples as explained before.

Table 2. Tests configurations

	Vertical Range	Horizontal Range	Size(s)	Positive Images	Negative Images
Case 1	-5˚ - 55˚	45˚ - 135˚	Single	28,665	116,260
Case 2	5˚ - 45˚	45˚ - 135˚	Single	20,475	83,500
Case 3	5˚ - 45˚	45˚ - 135˚	Three	61,425	247,300

The results achieved for detecting the dipper shows that the precision and accuracy of the detections are reduced by decreasing the latitude angles and increasing the number of sizes of the dipper (Table 3). That could happen based on the fact that the shape of the dipper is somehow symmetric; therefore, including multiple sizes of the dipper during the training phase led to the detection of other objects in the scene which look similar to the dipper.

Table 3. Results of experiment on static images

		Precision (%)	Recall (%)	Accuracy (%)
Dipper Detector	Case 1	54	93	52
	Case 2	40	100	40
	Case 3	10	100	10
Boom Detector	Case 1	64	100	64
	Case 2	80	98	78
	Case 3	86	98	84
Body Detector	Case 1	29	100	29
	Case 2	32	100	32
	Case 3	56	100	56

On the other hand, for the boom and the body, the results are opposite, and the precision and accuracy are increased by reducing the latitude angles and increasing the number of the sizes. This could be explained by the fact that the body (and specially the boom) looks unique and not symmetric; therefore, for cases two and three, the trained detectors have more chances to detect their target correctly. Regarding the recall, since the target objects are assumed to be always available in the scenes, there is no possibility for the true negatives, and the recall values have an opposite behavior compared to the precision and accuracy.

3.2 Experiment on Video Frames

This test is done on a construction site in Vancouver. The best detectors from Section 3.1 are selected to apply on the video acquired from the construction project. The detectors used in case one, case two, and case three for the dipper, boom, and body, respectively, are considered in this experiment.

The results in Table 3 show that there is no false negative detection. Also, there is no true negative since the excavator is always available in all video frames. The detectors perform well by achieving accuracies of 95% and 97% for the dipper and boom, respectively. However, the accuracy and precision of the body's detector are 48%, which is not satisfactory and needs to be further investigated.

Table 4. Results of experiment on video frames

	Precision (%)	**Recall (%)**	**Accuracy (%)**
Dipper	95	100	95
Boom	97	100	97
Body	48	100	48

4 CONCLUSIONS AND FUTURE WORK

In this paper, the proposed equipment parts' detection method was discussed. Before starting the pose estimation algorithm, it was necessary to create a comprehensive image database for the target equipment and its parts. Furthermore,

the database was used for training multiple CV detectors. The process of generating and annotating the images were performed using the method proposed in Soltani et al. (2015). The equipment parts' detectors were applied on the video/images from construction sites to search for the parts. Two experiments were done on the static images and video frames to validate the performance and capability of the proposed method. The results showed that reliable accuracies were achieved by applying the proposed method on the video frames. Moreover, the performance of High Performance Computing (HPC) on generating and annotating the image database showed that using more processors could dramatically reduce the required time.

REFERENCES

Azar, E. R., & McCabe, B. (2012). Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos. *Automation in construction, 24*, 194-202.

Azar, E. R., & McCabe, B. Y. (2013). A Visual Sensing Approach to Estimate Material Hauling Cycles in Heavy Construction and Surface Mining Jobsites. *The International Symposium on Automation and Robotics in Construction and Mining (ISARC 2013)*.

Azar, E. R., Dickinson, S., & 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.

Azar, E., Feng, C., & 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*, 213-229.

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.

Gong, J., & Caldas, C. (2010). Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations. *Journal of Computing in Civil Engineering, 24*(3), 252–263.

Gong, J., Caldas, C. H., & Gordon, C. (2011). Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. *Advanced Engineering Informatics, 25*(4), 771-782.

Olson, E. (2011). AprilTag: A robust and flexible visual fiducial system. *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA'11)*, pp. 3400-3407.

Renz, J., & Nebel, B. (2007). Qualitative spatial reasoning using constraint calculi. In *In Handbook of spatial logics* (pp. 161-215). Springer Netherlands.

Soltani, M. M., Zhu, Z., & Hammad, A. (2015). Developing Automated Annotation for Visual Recognition of Construction Resources. *2nd International Conference on Civil and Building Engineering Informatics (ICCBEI 2015)*. Tokyo, Japan.

Yang, J., Vela, P. A., Teizer, J., & Shi, Z. K. (2011). Vision-based crane tracking for understanding construction activity. *In Proceeding of ASCE Int. Workshop on Computing in Civil Engineering*, (pp. 258–265). Miami, FL .