This paper is published as:

Rezazadeh Azar, E, Dickinson, S, McCabe, BY, 2013, "Server-Customer Interaction Tracker (SCIT): A Computer Vision-Based System to Estimate Dirt-Loading Cycles" ASCE Journal of Construction Engineering & Management. 139(7):785-794

# Server-Customer Interaction Tracker (SCIT): A Computer Vision-Based System to Estimate Dirt-Loading Cycles

Ehsan Rezazadeh Azar[1], Sven Dickinson[2], and Brenda McCabe[3]

## Abstract

Real-time monitoring of the heavy equipment can help practitioners improve machine-intensive and cyclic earthmoving operations. It can also provide reliable data for future planning. Surface earthmoving job sites are among the best candidates for vision-based systems due to relatively clear sightlines and recognizable equipment. Several cutting edge computer vision algorithms are integrated with spatiotemporal information, and background knowledge to develop a framework, called server-customer interaction tracker (SCIT), which recognizes and measures the dirt loading cycles. The SCIT system detects dirt loading plants, including excavator and dump trucks, tracks them, and then uses captured spatiotemporal data to recognize loading cycles. A novel hybrid tracking algorithm is developed for the SCIT system to track dump trucks under visually noisy conditions of loading zones. The developed framework was evaluated using videos taken under various conditions. The SCIT system with novel hybrid tracking engine demonstrated reliable performance as the comparison of the machine-generated and ground truth data showed high accuracy.

**CE Database subject headings:** Construction management; automatic identification systems; data collection; imaging techniques; earthmoving; construction equipment.

## Introduction

Earthmoving operations are a major component of certain types of construction and mining industries, including heavy civil operations, sand and gravel pits, rock quarries, as well as surface mining for minerals, ores, and oil sands. These all depend on heavy equipment and have a repetitive character; slight reductions in cycle durations may result in considerable improvements in productivity, cost savings, and reductions in carbon emissions. For this reason, timely and reliable data are critical.

Manual real-time monitoring where workers watch the operations are costly, tedious, and error-prone. As a result, some sensing technologies, such as global positioning system (GPS) and radio frequency identification (RFID) have been used to monitor earthwork machines and to provide real-time locational data. GPS can calculate the three-dimensional

[1] Ph.D. Candidate, Department of Civil Engineering, University of Toronto, 35 St. George St., Toronto, ON. M5S1A4, Canada, e.rezazadehazar@mail.utoronto.ca

[2] Professor, Department of Computer Science, University of Toronto, 6 King's College Rd., Toronto, ON. M5S3G4, Canada, sven@cs.toronto.edu

[3] Associate Professor, Department of Civil Engineering, University of Toronto, 35 St. George St., Toronto, ON. M5S1A4, Canada, brenda.mccabe@utoronto.ca

location of the plant, and then a central processing unit analyzes the spatiotemporal pattern of the machine to recognize the action and therefore the productivity. However, there are a number of shortcomings associated with these devices. First, a receiver antenna needs to be installed on the equipment, which may be difficult for rented plants. Second, the acquired data is limited to time and location which sometimes makes it difficult to distinguish between productive and non-value added movements. Finally, these spatiotemporal records cannot help in the review process to find the events that caused abnormal cycles.

Computer vision-based methods are another possible solution to monitor earthmoving activities if clear sightlines can be selected and earthmoving machines are fairly recognizable at that distance. Different approaches are available to monitor earthmoving equipment using computer vision. The first approach is to develop software to analyze the visual capture from common electronic tools, such as video cameras. Recent wireless cameras and high-capacity storage devices are available at relatively low-cost, therefore it has become a common practice to monitor construction sites by CCTV systems (Zou and Kim 2007; Gong and Caldas 2011). Although this is a cheap alternative, normal cameras provide only 2D projections of the real world, thereby limiting the analysis to 2D. The second method is to employ other sensing devices together with video cameras to obtain more data from the scene. LASER range-finder, infrared cameras, and the application of multiple cameras at the same scene can provide 3D coordinates, but they are expensive and must remain accurately calibrated to work properly.

A novel pipeline framework developed in this research, called server-customer interaction tracker (SCIT), combines image and video processing modules, spatiotemporal reasoning, and expert logic. SCIT is used to detect the machines involved in loading actions, track them, recognize their interactions, and estimate the cycle times. Two different frameworks have been implemented using the same object recognition and action interpretation modules but different tracking engines: one has a classic tracking engine while the other uses an innovative framework to track the loading dump truck.

This research has two main contributions. First, it introduces a novel tracking algorithm to track dump trucks under moderate occlusion. Secondly, it proposes an action recognition framework to estimate dirt loading cycles in actual job site conditions.

The structure of the paper is as follows. First, it describes related advancements in construction research and highlights the motivations of this research. Then the modules and the structure of the SCIT framework are explained. Next, test results are provided for both systems using several test videos captured in different construction sites under varying visual conditions. Finally, the potential applications and shortcomings of the developed prototype are discussed.

## Literature Review

The earthmoving sector has a longer history of the application of automated monitoring systems than other segments of the construction industry (Navon 2005). Due to the permanent and manufacturing nature of surface mining operations, faster technology adoption to locate and dispatch the large equipment fleet has been encouraged. Sensing technologies were gradually employed in heavy construction projects due to common equipment used in mining and heavy construction.

GPS antennas have been the main tool to track machines in construction projects and then measure their productivity such as grading and leveling (Navon and Shpatnisky 2005; Navon et al. 2004) and asphalt paving (Navon and Shpatnisky 2005; Peyret et al. 2000). Manual preparation, indirect data interpretation, and the intrusive character of the GPS receivers are the main shortcomings of this system.

RFID tags were also used to estimate the loading, hauling, and dumping times of the dump trucks. Fixed readers installed in entrance gates of the loading and dumping sites record the entrance and exit of RFID tags attached to dump trucks. The time differences are considered as loading, traveling, and dumping cycle times (Montaser and Moselhi 2012). Implementation of this system is cumbersome in linear projects (e.g. highway construction). Further, it can not confirm whether the truck is actually loaded.

Computer vision techniques are relatively new data collection tools that have significant potential to monitor earthmoving equipment. Several advances have been achieved in construction for video processing including object tracking (Brilakis et al. 2011; Park et al. 2011) and object recognition (Rezazadeh Azar and McCabe 2012; Chi and Caldas 2011; Jog et al. 2011; Rezazadeh Azar and McCabe 2011), and productivity measurement including labour (Peddi et al. 2009; Weerasinghe and Ruwanpura 2009), concrete pouring (Gong and Caldas 2011), and earthmoving equipment (Gong and Caldas 2011). However, the research has a long way to go before a reliable and automated system can be implemented in the construction industry. This is due in part because computer vision is itself a continually evolving field of science. At the present time, existing algorithms for object recognition, tracking, and segmentation can fail under certain conditions, particularly in the visually noisy images and videos typical of a construction site. Further, the construction industry is resistant to new technologies (Navon and Sacks 2007), unless they are shown to reduce costs, be easy to use, and decrease the amount of non-productive time required from their staff. Despite the significant progress made to be able to use computer vision techniques in construction, much more research is needed as most developments to date have limited applicability, operate only under specific conditions, and require a high level of human intervention.

For example, color-based approaches to monitor earthmoving plants are not invariant against occlusion, lighting conditions, and the existence of other similarly coloured objects (Zou and Kim 2007). A vision-based system was developed to measure working cycles of a mini loader (Gong and Caldas 2011). It uses background subtraction to isolate moving equipment, and then classifies it using Bayes or neural network algorithms (Chi and Caldas 2011). This framework however, has difficulty in processing unfamiliar moving entities which typically appear in job sites. Detecting smaller tools that are dedicated to one activity type, such as a concrete hopper, also posed challenges as they required manual defining of the work zone (Gong and Caldas 2011, Almassi and McCabe 2008).

Our research intends to close the practicability gap between vision-based systems and earthmoving productivity measurement processes, where it can recognize and estimate dirt loading cycles under broader visual conditions such as different viewpoints and with the presence of multiple types of construction equipment. In addition, our system requires limited human intervention in the initial setup of the camera viewpoint.

## SCIT Modules

Both of the developed SCIT systems use the same object recognition techniques and action interpretation modules, but they have different trackers. The following subsections describe these modules.

### *Object recognition*

Details of the background effort for the object recognition module of this research to detect dump trucks is described in Rezazadeh Azar and McCabe (2011) where the Histogram of Oriented Gradients (HOG) (Dalal and Triggs 2005) algorithm used to detect dump trucks from eight viewpoints. Unfortunately, this method is computationally intensive and the runtimes were too high for real-time purposes. This is because the HOG object recognition method is rather brute force; and its classifier window searches for the target object in every location and scale of the image. For example, it takes about 26 seconds to scan a 640x480 frame for all eight orientations on a 2.93 GHz dual core CPU, as shown in Table 1.

Parallel implementation of the HOG algorithm using a Graphics Processing Unit (GPU) can accelerate the standard sequential code by over 67x (Prisacariu and Reid 2009). The parallel computing platform and programming model (CUDA) technology developed by NVIDIA is used for this research (NVIDIA 2012). This new generation of the GPUs have hundreds of cores, allowing them process thousands of threads in parallel and enable non-uniform access to memory (NUMA). In this framework, the host CPU first loads the frame and copies it to GPU memory. The GPU processes all of the scales and windows of the image and returns calculated support vector machines (SVM) scores of each search window to the host CPU. The host CPU formats the results, which include the score and position of each window, and finally performs the non-maximal suppression to fuse the detected boxes. The reason to process fusion operations on CPU is that they require a lot of connections to RAM.

The runtimes for scanning the same eight views with the same CPU (2.93 GHz dual core) and a GeForce GT 440 GPU with 2.1 compute capability were reduced drastically (see Table 1). Therefore, it is now possible to use the standalone HOG method as the truck recognition engine of the system, with an acceptable detection rate, and maintain the real-time video stream.

**Table 1. Runtimes of the HOG recognition process for eight viewpoints of a dump truck**

| Detector / Image size | CPU Dual core 2.93 GHz | | GPU NVIDIA GeForce GT440 |
|---|---|---|---|
| | HOG (Sec) | H&H (Sec) | HOG (Sec) |
| 640x480 | 26 | 1.2~2.6 | 1.07 |
| 1024x768 | 69 | 2.5~5.3 | 2.8 |
| 1920x1080 | 186 | 6.9~14 | 7.6 |
| 2592x1944 | 455 | 19.6~25.9 | 18.8 |

Unlike the rigid figures of dump trucks, the articulated features of hydraulic excavators make them a more difficult recognition target. A recognition algorithm was developed that uses a part-based approach and spatiotemporal reasoning to detect an excavator in consecutive frames of a video (Rezazadeh Azar and McCabe 2012). The excavator recognition module in SCIT has been also implemented using GPU which the recognition process of 640x480 frames has been reduced from 6.5-7.0 seconds to 0.35-1.2 seconds.

### *Mean-shift tracking*

In comparative studies of tracking algorithms applied to the harsh and visually noisy construction environment, researchers agree that the Mean-shift algorithm is reliable for tracking equipment (Gong and Caldas 2011; Park et al. 2011), and the addition of the Kalman filter and Particle filter can stabilize its performance (Gong and Caldas 2011). Mean-shift tracking is a Kernel-based algorithm that searches for a local peak in the density distribution of a dataset, and ignores the outliers far from the maxima (Comaniciu et al. 2003). As a result, the first framework developed here used a modified version of the Mean-shift algorithm, called continuously adaptive Mean-shift or Camshift (Bradski 1998), as the main tracking engine. Mean-shift and Camshift trackers can track different feature types; application of hue, saturation, and value (HSV) color features is the most common approach. HSV color histogram and intensity response of the HOG detectors were used as tracking features for this research. In the second approach, the dense greyscale image of the HOG detector response is provided for the Mean-shift tracker. The maximum response is colored with the highest intensity and the rest of the responses are normalized based on their HOG detection score. The Mean-shift method however, has a limitation: since this algorithm searches for local maxima, the tracking blob may expand or shift to a nearby object of similar color or to similar textures that the algorithm uses. This issue, along with other problems we encountered, are described further in the discussion section.

### *Hybrid tracking*

Because existing tracking methods have difficulty performing well when applied to construction videos (Gong and Caldas 2011; Park et al. 2011), a new hybrid tracking technique was developed by the authors to overcome the shortcomings of the Mean-shift method. The hybrid tracking method is inspired by a new recognition-based tracking and activity interpretation framework (Barbu et al. 2012). They employed a latent SVM detection method (Felzenszwalb et al. 2010) with lowered thresholds to produce tracking candidates, and also applied a feature tracker (Tomasi and Kanade 1991) to project each detection five frames forward to compensate for false negatives of the raw detector. Then a dynamic-programming algorithm (Viterbi 1971) selects a temporally coherent set of detections for tracking.

Our hybrid technique incorporates the HOG algorithm to help track identified equipment. This is achievable because equipment profiles do not change drastically between time steps. After recognizing that a truck is being loaded, the system continues to search for trucks, but in an optimized manner using much shorter time intervals. In these settings, the recognition module searches for only three orientations every two seconds, i.e., the initial orientation of the target dump truck and the two adjacent viewpoints, which takes about 0.39 seconds with the same GPU and processor. For example, if the target truck was in a side-left orientation, the framework only scans for front-left, side-left, and rear-left viewpoints. This way, changes in the trajectory of the machine are caught quickly, but the computational effort required to check all eight orientations is not necessary using a priori knowledge. In addition, the thresholds are decreased to avoid false negatives; however, the rate of false alarms increases as well. Each detection creates a bounding box to mark the location of the potential target.

Two issues exist with the pure recognition-based tracking. First, even lowering the thresholds cannot guarantee continuous detection of the equipment, which can result in

losing the target. Second, there were several instances in the test videos where a second truck entered the frame and waited in the loading zone with a similar orientation as the truck being loaded. This often misled the recognition-based tracking algorithm, as did nearby false positives.

Thus, a tracking tool was added to the framework to artificially generate new bounding boxes in new frames to eliminate the risk of losing the truck, and also keep track of the actual target. In this approach, the center point of the loading truck is tracked by the KLT feature tracker (Tomasi and Kanade 1991) to project that bounding box to the next scanning frame. The KLT method is a differential method to estimate the optical flow, which is based on three assumptions: 1) brightness constancy, 2) temporal persistence, and 3) spatial coherence. The KLT optical flow estimator projects a rectangle in addition to the true positive and false alarm windows generated by the recognition engine in every new frame. A simple disjoint-set data structure algorithm is then used to partition the detection that is temporally coherent with the projected rectangle, yielding a new bounding box and eliminating other boxes. Two boxes are considered to be in the same subset if their bounding regions overlap. All of the distances between x and y elements of the matching corners should be less than the minimum average of the width and height of the boxes times a threshold to group two rectangles (Viola and Jones 2001). The corners of the final box are the average of the corners of the projected rectangle and the overlapping detection. If there is not temporally coherent detection, the projected box will be taken as the final rectangle.

The flowchart and the visual sequence of the entire process are presented in Figure 1 and Figure 2, respectively. Figure 2a shows the truck of interest (the next section explains how to select the loading truck among others), and frames b and c depict the result of the object recognition and projected box by KLT tracking two seconds later. The final fusion result of the algorithm is presented in Figure 2d. After creation of the new box, the center of that rectangle becomes a feature for the KLT tracker. The KLT feature tracker is very sensitive to any object passing in front of the tracking features, such as workers or the bucket of an excavator. Even the shadow of the bucket could distract the tracker; however, the hybrid nature of this novel tracker means that the continuous HOG object recognition in short time intervals prevents equipment from being lost, improving the performance of this hybrid tracker.

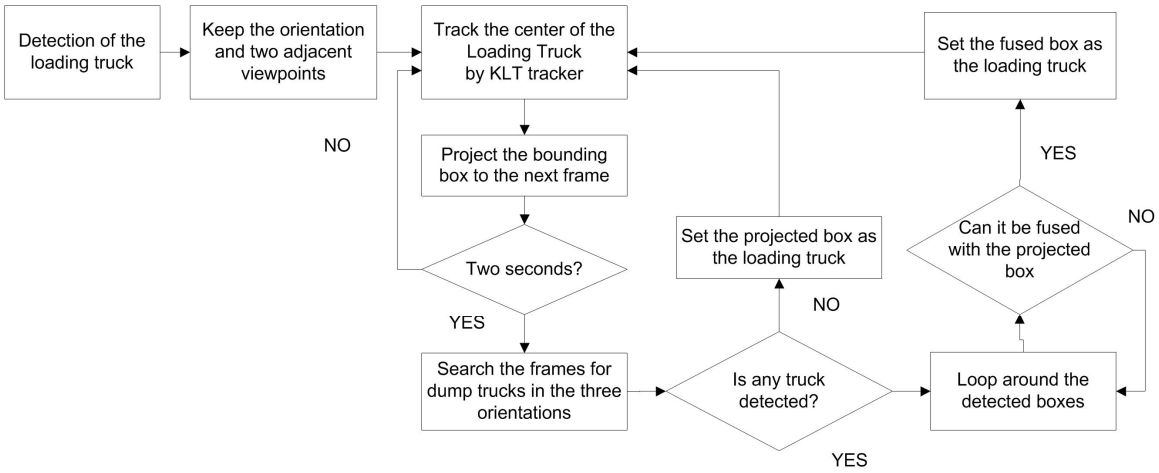


**Figure 1. Flowchart of the hybrid tracking algorithm**

### *Action recognition module*

There are two main activity recognition approaches in the computer vision community: logical reasoning and artificial intelligence plan recognition algorithms. In logic-based methods, a set of consistent logical constraints is developed where all must be met to confirm the action. In artificial intelligence approaches, however, probability theory and machine learning models, such as Hidden Markov Models and support vector machines, are used to interpret the events and allow for uncertainty. It was argued that any plan recognition system should include some theory of uncertainty (Charniak and Goldman 1993) to find the most probable plan between two or more feasible candidates. In addition, artificial intelligence task recognition algorithms can continually learn the new unknown situations.

Every detected dump truck and excavator is defined by the system with its 2D coordinates, dimensions of the bounding box that represents the detected equipment, and its orientation. To identify a loading action, a dump truck should be within range of the excavator's boom and be in the proper orientation. Thus, the configuration and the distance of the equipment are two key data in recognition of the action. Together with time, these spatiotemporal data are used in SCIT to interpret the interaction of the detected machines.

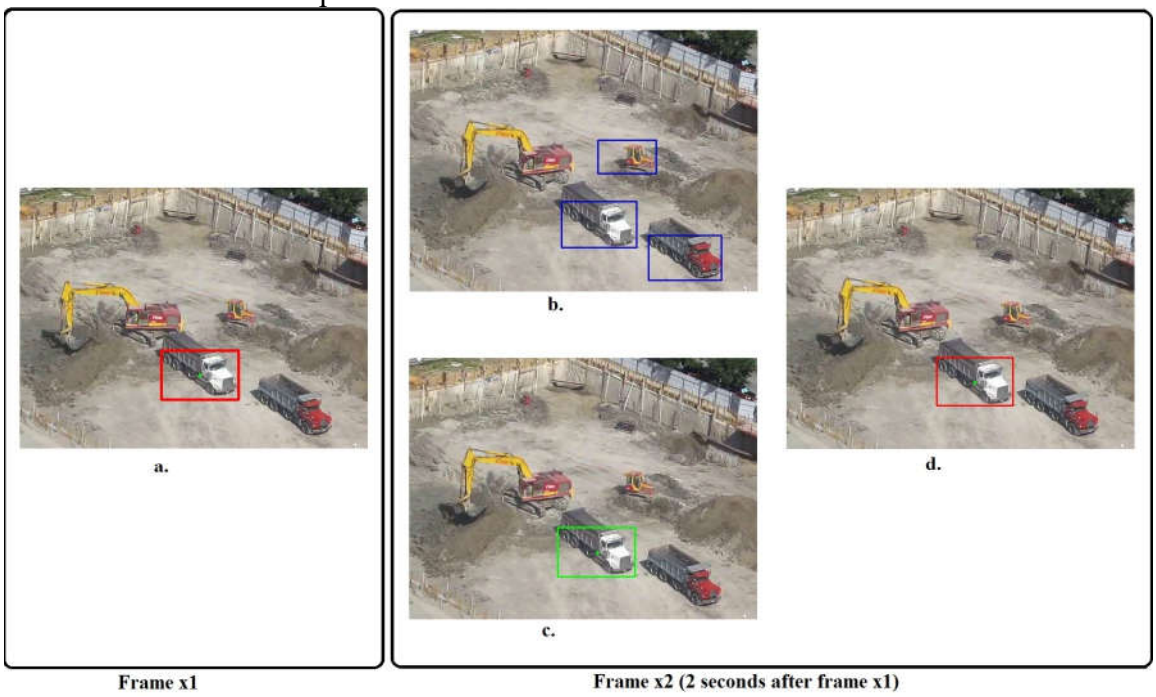


**Figure 2. a: detected truck at frame x1; b: HOG recognition result with lowered thresholds for three viewpoints in frame x2; c: projected box of previous frame (frame x1) to frame x2 using KLT feature tracker; d: fusion of the rectangles in b and c**

A combination of the logical reasoning and a pattern recognition algorithm were used to develop a module which can recognize dirt loading activity. The first component of this algorithm is a logical reasoning framework that checks equipment orientations for loading. For example, a right facing boom and a side-left dump truck located in the right of the excavator will never result in loading. Since this logical reasoning only intends to filter candidates, inclusion of uncertainty is not an issue. Table 2 presents the logical loading configurations based on the location of the excavator and orientation of dump trucks.

**Table 2: Possible loading configurations**

| Excavator / Dump truck | Left side of the dump truck | Right side of the dump truck |
|---|---|---|
| Front | ✓ | ✓ |
| Front-left | | ✓ |
| Front-right | ✓ | |
| Side-left | | ✓ |
| Side-right | ✓ | |
| Rear | ✓ | ✓ |
| Rear-left | | ✓ |
| Rear-right | ✓ | |

Dump trucks that pass this phase will be sent to the second stage, which examines the distance and size ratio of the server and customer. The corner of the excavator boom's bounding box closest to the hinged support of the boom is set as the base point. For example, if the boom is right-sided, the base point would be the bottom left corner of the bounding rectangle. The algorithm measures the distances between the base point of the excavator and four corners of the dump truck, and then these distances are divided by the width of the excavator bounding box to incorporate the size factor. Figure 3 illustrates these distances. The resulting numbers form a vector with four elements. Several positive and negative sample vectors were collected to train a linear support vector machines (Cortes and Vapnik 1995) under supervised learning as the second filter of the action recognition process. Seven videos with total duration of fifty one minutes were used to train the action recognition classifier. The bounding boxes of the dump trucks appearing in the videos were manually labeled as if they are being loaded by the existing excavator. This manual labeling produced 514 positive and 828 negative samples to train the classifier. Open source SVM-light software (Joachims 1999) trained the classifier.

**Figure 3: Distances between the corners of trucks and the base point in both left and right configurations**

This classifier checks the dump trucks that passed the first phase. The distances between the base point of the excavator and four corners of the available dump trucks are computed and then divided by the excavator's width to produce a vector with four elements, which was then classified by the trained SVM classifier. The scalar product of the classifier and test vectors produces a score and the objects with scores greater than a threshold are accepted. If two or more dump trucks pass the classification stage, the system will pick the machine with the highest classification score as a loading truck.
In addition to action recognition, these constraints enable the system to ignore the trucks and false positives which do not meet the constraints, and allow the truck recognition thresholds to be lowered to result in fewer false negatives.

## Structure of SCIT Systems

As mentioned before, the SCIT system was developed using two different tracking engines and this section explains the structure of both SCIT frameworks. They were implemented using open source OpenCV 2.3.1 library (OpenCV 2011) in Visual C++ express 2010 environment.
In this approach, the system searches for an excavator. Once found (Figure 4.a), it sends the detected bounding box to the Mean-shift tracking engine (Figure 4.b). The current prototype of SCIT stops searching at one excavator, but it is possible to modify the framework to process videos with two or more loading units. In addition to tracking the excavator, the system also starts to scan for dump trucks in predetermined time intervals (see Figure 4.c). Since it takes about 1.07 seconds to scan a 640x480 pixel frame for all eight orientations, the system can be set to search for dump trucks in any time interval greater than 1.07 seconds. In this case, the recommended four second intervals were used (Rojas 2008). The action recognition module analyzes all of the detected dump trucks in each recognition interval to check if any of them pass the classification. As soon as the system confirms the loading action (see Figure 4.d), it will stop searching for dump trucks and pass the loading truck to the tracking module (see Figure 4.e, this image shows the Mean-shift tracker), define the loading zone, and record the start time of the loading cycle. The loading zone is defined as 1.25 times the truck length and 1.5 times the truck height (dark blue rectangles in frame d and e in Figure 4). This loading zone is defined relatively large due to minor movements of dump trucks during loading for better positioning, and to accommodate the relatively small spatial variations used by the tracking algorithms thereby reducing the risk of early termination of the tracking. As shown in frame d and e of Figure 4, the center of the loading zone is not the same as the center of the detected truck and is shifted toward the hydraulic excavator. This formation handles the slight truck movements, which are mostly backward, and reduces the difference between the actual and SCIT finish times.
The tracking module continues tracking the truck until the center of the tracking blob exits the loading zone. At that moment, it records the finish time, terminates tracking of the dump truck, deletes the loading zone (see Figure 4.f), and begins to search for new dump trucks. The entire flowchart of this frame work is depicted in Figure 5.

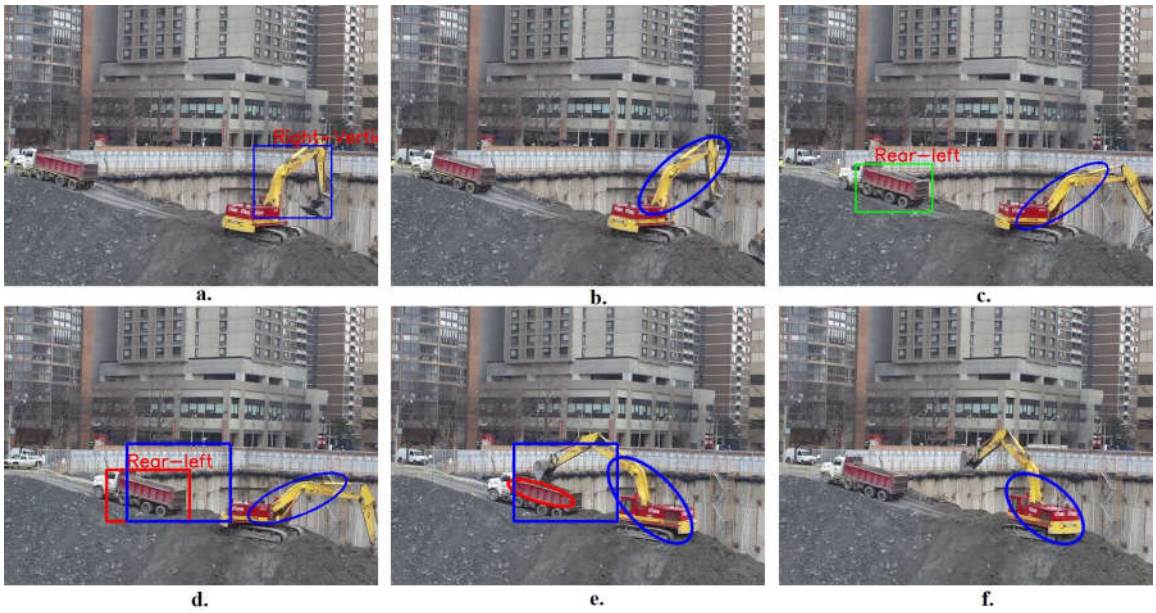


**Figure 4. a: Detection of the excavator; b: tracking the excavator; c: detection of a truck that does not meet loading criteria; d: detection of the loading truck; e: tracking of the both equipment; f: truck leaves the zone and tracking of the truck terminates**

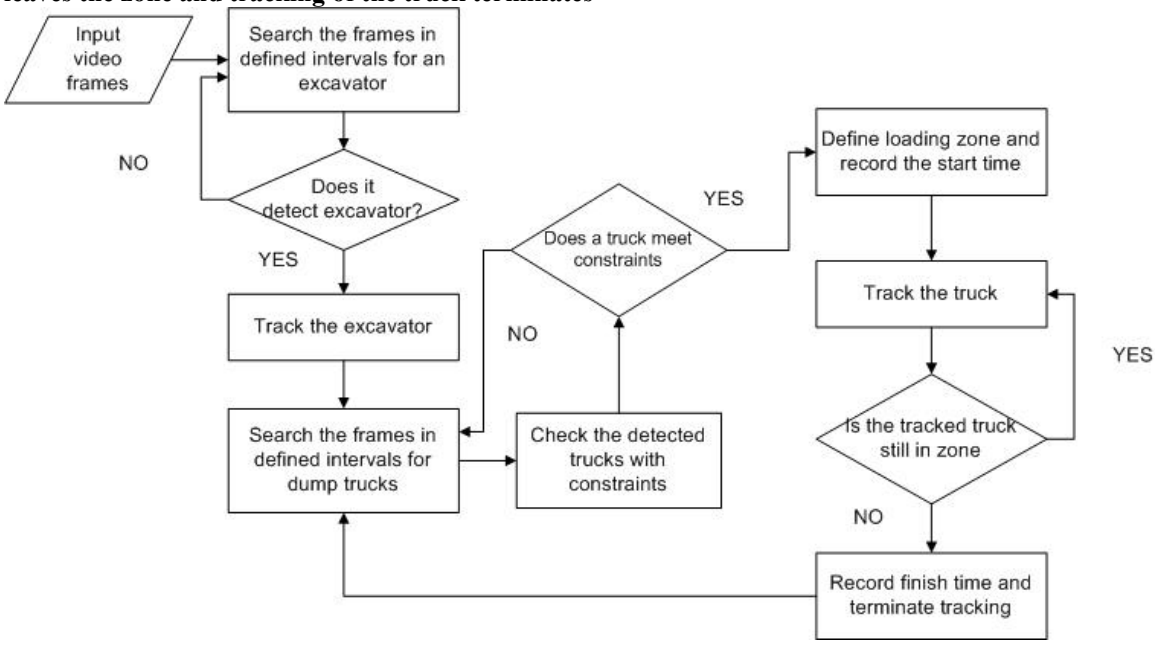


**Figure 5: Flowchart of the SCIT systems**

## Experimental Results

To test and compare the two developed methods, eighteen videos with total duration of two hours and twenty seven minutes were captured of excavation activities at two condominium complexes in downtown Toronto, Ontario. None of these videos were used in training stage. The equipment of these two projects had very similar productivity rates, which resulted in a homogenous productivity dataset. Two types of hydraulic excavators

(Caterpillar 245B and Caterpillar 345D), and several makes of urban dump trucks with similar hauling capacity, such as Mack, Sterling, Volvo, and Kenworth appeared in the videos. These videos were recorded by one of the authors during eight site visits under different lighting conditions in all four seasons, from ground and elevated angles with two different makes of digital cameras. Figure 6 shows some of these views.

The excavators had typical construction colors, but the urban dump trucks were painted in a wide range of colors. The SCIT systems processed the test videos with varied action recognition thresholds and the results are presented together with manual observation as ground truth in Table 3. This table shows the statistics of true positives, false alarms, and average cycle times of true positive cycles. In the manual observation, the loading time starts when a dump truck completely stops in front of the excavator; the action ends as the truck starts moving out of the zone. The tests with smaller ID numbers have lower action recognition thresholds, and the threshold rises as the test number increases.


**Figure 6. Some of the dirt loading views**

**Table 3: Results of the experiments on test videos**

| | Number of Cycles | | | Average true positive cycle time Seconds |
|---|---|---|---|---|
| | True positive cycles | False negative cycles | False positive cycles | |
| Manual | 55 | 0 | 0 | 101.87 |
| SCIT with hybrid, test 1 | 53 | 2 | 4 | 106.49 |
| SCIT with hybrid, test 2 | 54 | 1 | 2 | 106.43 |
| SCIT with hybrid, test 3 | 54 | 1 | 2 | 105.93 |
| SCIT with hybrid, test 4 | 51 | 4 | 1 | 105.08 |

| | | | | |
|---|---|---|---|---|
| SCIT with Mean-shift , test 3 | 30 | 1 | 2 | 105.00 |

As shown in Table 3, the best performance was SCIT with hybrid, test 3. The threshold of test 3 is optimal as it had the highest true positive cycles along with test 2, and its deviation was less than test 2. Only one test round was carried out with the Mean-shift algorithm (one with HSV colors and one with HOG response) using the optimal threshold found in the Hybrid 3 test, but the performance was substandard, so the other tests were aborted.

## Discussion

Lowering the threshold increases the risk of false positives e.g. accepting a false positive object or accepting a waiting truck as the one being loaded. In addition, lower thresholds may produce longer cycle times as they detect the loading truck before it completely positions for loading. The results reveal that increasing the thresholds improved the average time. However, it improves the performance up to a certain level (hybrid test 3 in Table 3) and afterwards causes some missing cycles as the SVM classifier does not accept the actual loading truck. For instance, the test 4, which had the highest threshold, missed three more cycles than the test 3.

All of the processed videos were studied manually to identify the root of the errors. Both systems failed to recognize one loading cycle due to occlusion where the loading truck was entirely masked by another truck. This case is difficult even for a human to track (Figure 7a, b, and c). Both of the frameworks detected and tracked the foreground machine (green truck) instead of the real loading truck (dark blue) and produced wrong outcomes. This only occurred in a video captured from ground level; videos with overlooking views did not have trucks mostly masked by another. Frames d, e, and f in Figure 7 show the same work zone with the same configuration of the equipment that were captured only a few minutes later from a higher angle. In this case, the SCIT with hybrid tracker correctly detected and tracked the loading truck. These results show the importance of the proper viewpoint of the camera for robust outcomes. The camera can be installed on nearby buildings, tower cranes, peaks of slopes, or temporary posts; however, some construction sites do not have this luxury where there are no overlooking points around, or they are not accessible. Application of multiple cameras is another possible solution for such cases, but this is beyond the scope of this research.

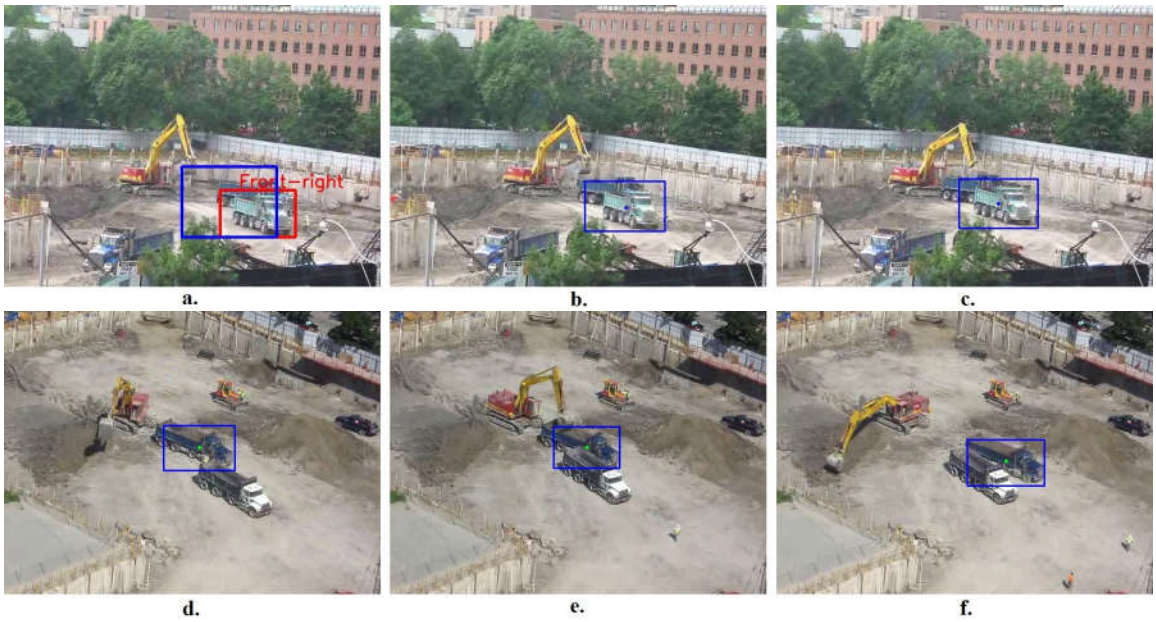


**Figure 7. Recognition and tracking of the wrong loading truck due to severe occlusion**

The SCIT with the Mean-shift algorithm using color features had serious difficulty tracking dirty gray, white, or black dump trucks as their color histogram was similar to the background color (frames a and f in Figure 6). In several cases, it missed the target, thereby thinking that the loading activity was complete. This resulted in early termination of the loading clock, resetting of the cycle upon the next detection of the same truck, and therefore inaccurate productivity data. Since the HOG method is invariant to color, the Mean-shift with HOG response and the hybrid algorithms could successfully process those cycles. In addition, the Mean-shift method with color histograms failed in some cases where the tracking blob moved or expanded from the target to a nearby truck of the same color (frames a to c in Figure 8), again resulting in incorrect results. The same issue occurred in Mean-shift with HOG response regardless of the machine color, because both dump trucks had similar orientations resulting in high detection scores. SCIT with the hybrid tracker flawlessly processed all of the mentioned cases (frames d to f in Figure 8).

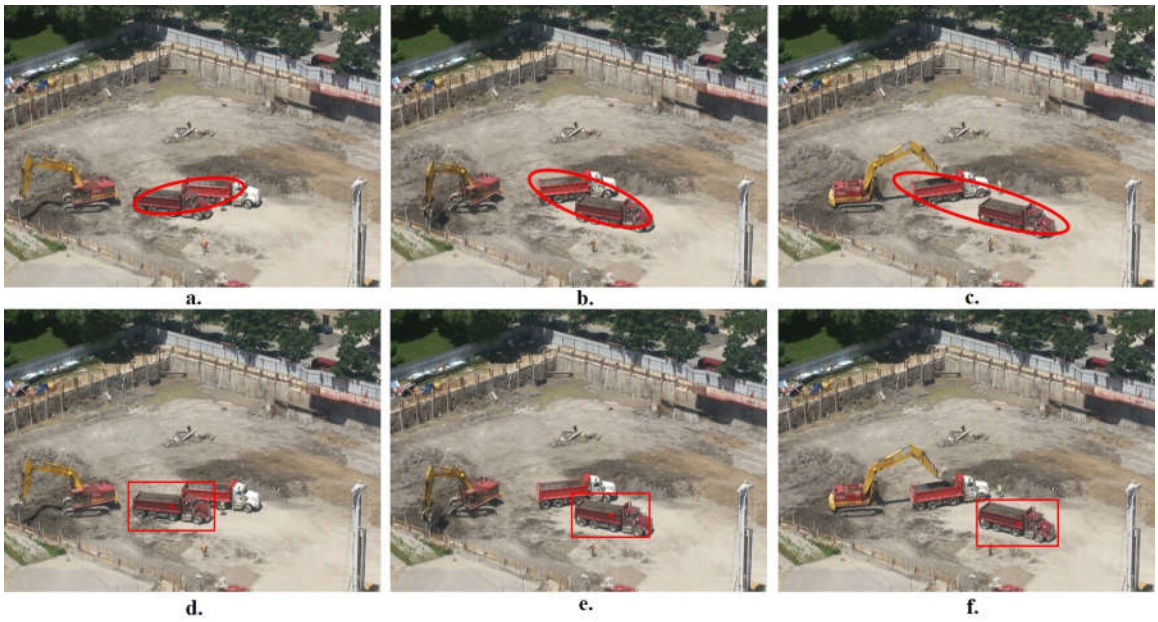


**Figure 8. a through c: mixture and wrong shifting of the tracking blob to nearby same colored truck, d to f: hybrid algorithm correctly tracked the target in the same scenario**

Both methods, regardless of their tracking engine, had some variations in recording the start and finish times, which resulted in different average cycle times (Table 3). In reviewing the data, the errors fell into three main causes:

- The videos were scanned for new dump trucks every four seconds, so 0 to 4 second variations of the activity start times compared to manual observation are inevitable;
- SCIT was slow to find the loading truck although it was in place when it was scanning for it;
- It takes a few seconds after the truck begins to pull away for the center of the tracking blob to leave the loading region. This resulted in variations in recording the activity finish times in SCIT whereas the human observer instantly recorded the finish time when the loading truck started to leave.

## Practical Applications

There are two main applications for the automated productivity measurement of dirt loading cycles: production confirmation and productivity improvement.

The system can count the number of trips made by an earthmoving subcontractor, thereby confirming the work achieved for payment. It is also possible to approximate the quantity of earth moved using the number of trips and the standard capacity of dump trucks. These data are currently handled by foremen, who are also responsible for directing trucks in the loading zone. This is one of the most hazardous areas in construction sites due to slewing excavators, and the forward and reverse movement of dump trucks in a confined area (Edwards and Nicholas 2002). SCIT will eliminate the distracting recording task and help the foremen focus on site safety.

Activity duration data can be used to study productivity, find bottlenecks, and enhance ongoing operations. They can also be used in advanced analysis such as stochastic simulation for planning future activities (AbouRizk and Halpin 1992).

Table 4 provides the detailed results of test number 3. This table excludes the one false positive cycle and is based on the rest of 54 correctly detected cycles. These productivity data are grouped into the eight site visits because the productivity durations and visual conditions were different in each visit. The best case had 100% accuracy in average cycle times (site visit 6) and the worst was 86.3% (site visit 2). The overall accuracy of the average cycle times was 95% with the standard deviation of 4.5%. The accuracy of average cycle times is calculated as: 1-(software data-manual data)/manual data.

**Table 4. Detailed results of the dirt loading analysis**

| Site visit | No of cycles | Data Type | Waiting time | Loading time | Waiting | Loading | Accuracy | Average loading time | Average waiting time |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 17 | Manual | 0:16:05 | 0:28:54 | 35.75% | 64.25% | 98.04% | 0:01:42 | 0:00:57 |
| | | Software | 0:15:12 | 0:29:47 | 33.79% | 66.21% | | 0:01:45 | 0:00:54 |
| 2 | 6 | Manual | 0:04:17 | 0:10:12 | 29.57% | 70.43% | 90.45% | 0:01:42 | 0:00:43 |
| | | Software | 0:02:54 | 0:11:35 | 20.02% | 79.98% | | 0:01:56 | 0:00:29 |
| 3 | 3 | Manual | 0:06:14 | 0:03:55 | 61.41% | 38.59% | 95.73% | 0:01:18 | 0:02:05 |
| | | Software | 0:05:48 | 0:04:21 | 57.14% | 42.86% | | 0:01:27 | 0:01:56 |
| 4 | 2 | Manual | 0:02:28 | 0:04:32 | 35.24% | 64.76% | 97.62% | 0:02:16 | 0:01:14 |
| | | Software | 0:02:18 | 0:04:42 | 32.86% | 67.14% | | 0:02:21 | 0:01:09 |
| 5 | 3 | Manual | 0:02:25 | 0:05:42 | 29.77% | 70.23% | 95.69% | 0:01:54 | 0:00:48 |
| | | Software | 0:02:04 | 0:06:03 | 25.46% | 74.54% | | 0:02:01 | 0:00:41 |
| 6 | 5 | Manual | 0:04:16 | 0:06:11 | 40.83% | 59.17% | 99.52% | 0:01:14 | 0:00:51 |
| | | Software | 0:04:19 | 0:06:08 | 41.31% | 58.69% | | 0:01:14 | 0:00:52 |
| 7 | 6 | Manual | 0:04:53 | 0:10:27 | 31.85% | 68.15% | 97.17% | 0:01:44 | 0:00:49 |
| | | Software | 0:04:27 | 0:10:53 | 29.02% | 70.98% | | 0:01:49 | 0:00:45 |
| 8 | 12 | Manual | 0:12:36 | 0:21:25 | 37.04% | 62.96% | 98.73% | 0:01:47 | 0:01:03 |
| | | Software | 0:12:10 | 0:21:51 | 35.77% | 64.23% | | 0:01:49 | 0:01:01 |
| Overall | 54 | Manual | 0:53:14 | 1:31:18 | 36.83% | 63.17% | 97.21% | 0:01:41 | 0:00:59 |
| | | Software | 0:49:12 | 1:35:20 | 34.04% | 65.96% | | 0:01:46 | 0:00:55 |

These promising results demonstrate the practicability of the SCIT system to estimate dirt loading cycles, but they also highlight the constraints to employ this system. It is critical for the system to have a clear view of the scene and the loading trucks should be mostly

visible during the loading operation. Moreover, since the object recognition framework for excavator can currently detect only one excavator, the SCIT performance is restricted to one operating excavator and the job sites with more than one loading units require more cameras accordingly.

## Conclusion

Fifty-five years ago, Herbert Simon promised that in a foreseeable future, machines could think, learn, and create like a human (Simon and Newell 1958). We apparently are far from that goal even after half a century, although the technology is gradually moving in that direction.

In this research, we introduce a vision-based framework, named SCIT, which can recognize and estimate dirt loading cycles. The prototype software was developed using two different tracking methods and then evaluated using several test videos captured from two construction sites under different visual conditions. The SCIT with the novel hybrid tracking engine outperformed the SCIT with the well-known Mean-shift tracker. The results showed that this system could recognize and measure 98.2% of the loading cycles with 95% accuracy in durations; however, the system is vulnerable against harshly occluded target objects. From this issue, the proper location of the camera is seen to be a key factor for accurate results. In addition, the current version of the SCIT can only process videos with one operating excavator. The contribution of this research is to introduce a novel tracking method and an action recognition module to practically estimate dirt loading cycles. The future work will focus on the expansion the system to recognize and measure other types of earthmoving operations with multiple plants.

## References


AbouRizk, S.M., and Halpin, D.W. (1992). "Statistical Properties of Construction Duration Data." *Journal of Construction Engineering and Management*, 118(3): 525-544.

Almassi, A.N., and McCabe, B.Y. (2008). "Image Recognition and Automated Data Extraction in Construction." *Proc., Canadian Society of Civil Engineering Annual Conference*, Quebec, QC, Canada, Paper GC-568.

Barbu, A., Bridge, A., Burchill, Z., Coroian, D., Dickinson, S., Fidler, S., Michaux, A., Mussman, S., Narayanaswamy, S., Salvi, D., Schmidt, L., Shangguan, J., Siskind, J.M., Waggoner, J., Wang, S., Wei, J., Yin, Y., Zhang, Z. (2012). "Video In Sentences Out." *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina, CA.

Bradski, G.R. (1998). "Real time face and object tracking as a component of a perceptual user interface." *Proc., Applications of Computer Vision*, Princeton, NJ, USA. 214 – 219.

Brilakis, I., Park, M.W. and Jog, G. (2011). "Automated Vision Tracking of Project Related Entities", *J. of Advanced Engineering Informatics*, 25(4), 713-724.

Charniak, E., and Goldman. R.P. (1993). "A Bayesian model of plan recognition". *Artificial Intelligence*, 64, 53–79.

Chi, S., and Caldas, C.H. (2011). "Automated Object Identification Using Optical Video Cameras on Construction Sites." *Journal of Computer-Aided Civil and Infrastructure Engineering*, 26, 368–380.

Comaniciu. D., Ramesh, V., and Meer, P. (2003). “Kernel-Based Object Tracking.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(5), 564-577.
Cortes, C, and Vapnik, V. (1995). "Support-Vector Networks", *Machine Learning*, 20(3), 273-297.
Dalal, N., and Triggs. B. (2005). “Histograms of Oriented Gradients for Human Detection.” *Conference on Computer Vision and Pattern Recognition*, IEEE, San Diego, CA, USA, 2, 886 – 893.
Edwards, D.J., and Nicholas, J. (2002). “The state of health and safety in the UK construction industry with a focus on plant operators” *J. Structural Survey,* 20(2), 78-87.
Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D. (2010). “Object Detection with Discriminatively Trained Part Based Models.” *Journal of Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 32(9), 1627 – 1645.
Gong, J., and Caldas, C.H. (2011). “An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations.” *Automation in Construction*, in press.
Joachims. T. (1999). “Making large-scale SVM learning practical.” *Advances in Kernel Methods: Support Vector Learning*, B. Schlkopf, C. Burges, and A. Smola, The MIT Press, Cambridge, MA, USA.
Jog, G.M., Park, M.-W., and Brilakis, I. (2011). “Truck-Face Recognition using Semantic Texton Forests.” *Proc., 3rd International/9th Construction Specialty Conference*, CSCE, Ottawa, Canada, CN-87.
Montaser, A., and Moselhi, O. (2012). “RFID+ for Tracking Earthmoving Operations.” *Construction Research Congress*, West Lafayette, IN, USA, 1011-1020.
Navon, R., Goldschmidt, E., and Shpatnisky, Y. (2004). “A concept proving prototype of automated earthmoving control.” *Journal of Automation in Construction*, 13(2), 225–239.
Navon, R. (2005). “Automated project performance control of construction projects.” *Automation in Construction*, 14(4), 467– 476.
Navon, R., and Shpatnisky, Y. (2005). “Field Experiments in Automated Monitoring of Road Construction.” *Journal of Construction Engineering and Management*, 131(4), 487– 493.
Navon, R., and Sacks, R. (2007). “Assessing research issues in Automated Project Performance Control (APPC).” *Automation in Construction,* 16(4), 474–484.
NVIDIA (2012). “CUDA Parallel Programming.” <http://www.nvidia.com/object/cuda_home_new.html> (February 12, 2012).
OpenCv (2011). “The OpenCv Library.” < http://opencv.willowgarage.com/wiki/> (Dec. 10, 2010).
Park, M.-W., Makhmalbaf, A., and Brilakis, I. (2011). “Comparative study of vision tracking methods for tracking of construction site resources.” *Automation in Construction*, 20(7), 905-915.
Peddi, A., Huan, L., Bai, Y. and Kim, S. (2009). “Development of human pose analyzing algorithms for the determination of construction productivity in real-time.” *Building a sustainable future, Construction Research Congress*, Seattle, WA, USA, 1: 11-20.
Peyret, F. Betaille, D., and Hintzy G. (2000). “High-precision application of GPS in the field of real-time equipment positioning.” *Automation in Construction*, 9(3), 299-314.

Prisacariu, V., and Reid, I. (2009). "FastHOG - a realtime GPU implementation of HOG." Technical report, Department of Engineering Science, Oxford University, UK.

Rezazadeh Azar, E., and McCabe, B. (2011). "Automated Visual Recognition of Dump Trucks in Construction Videos." *Journal of Computing in Civil Engineering*, in press.

Rezazadeh Azar, E., and McCabe, B. (2012) "Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos." A*utomation in construction*, 24, 194-202.

Rojas, E.D. (2008). "Construction Productivity: A Practical Guide for Building and Electrical Contractors." *J. Ross Publishing*, Fort Lauderdale, Florida.

Simon, H., and Newell, A. (1958). "Heuristic problem solving: The next advance in operations research." *Operations Research* 6(1), 1-10.

Tomasi, C., and Kanade, T. (1991). "Detection and tracking of point features." *Technical Report CMU-CS-91-132*, Carnegie Mellon University.

Viola, P. and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features." *Proc., IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'01)*, IEEE, Kauai, HI, USA, 1, 1-9.

Viterbi, A. J. (1971). "Convolutional codes and their performance in communication systems." *IEEE Transactions on Communication*, 19(5),751–772.

Weerasinghe, I.P.T., and Ruwanpura, J.Y. (2009). "Automated data acquisition system to assess construction worker performance." *Building a sustainable future, Construction Research Congress*, Seattle, WA, USA, 1: 61-70.

Zou, J., and Kim, H. (2007). "Using Hue, Saturation, and Value Color Space for Hydraulic Excavator Idle Time Analysis." *Journal of Computing in Civil Engineering*, 21(4), 238-246.