Active Control of a Pan-Tilt-Zoom Camera for Vision-Based Monitoring of Equipment in Construction and Surface Mining Jobsites

E. Rezazadeh Azar[a]

[a] Department of Civil Engineering, Lakehead University, Canada
E-mail: eazar@lakeheadu.ca

Abstract –

Automated monitoring systems have proven to be effective in improving the productivity of equipment-intensive operations in the construction and mining sectors. Vision-based systems are the most recent methods employed to detect, track, and monitor construction equipment. The currently developed systems, however, analyze video frames captured by a stationary camera, which dramatically limits their coverage area and requires manual adjustment of the viewfinder. This research paper introduces methods to proactively steer a pan-tilt-zoom camera to localize, track, and identify objects of interest in construction jobsites. This automated camera control system uses a number of image and video processing algorithms to detect objects and estimate their trajectory and velocity, and then uses the extracted information to set the camera movement parameters, including direction and magnitude. The experimental results of this system showed promising performance for equipment monitoring in construction and mining jobsites.

Keywords –

Computer vision; equipment monitoring; pan-tilt-zoom camera; automated control

1 Introduction

There is a growing trend in the construction and mining industries to use cameras to monitor jobsites [1], [2], [3]. The resulting videos, however, are mainly processed manually, which is cumbersome, labor-intensive, and hinders the potential use of these rich resources. Therefore, there has been considerable interest in the construction research community to apply computer vision methods to automate video analysis processes. In particular, research efforts have focused on detection, localization, and tracking construction equipment for productivity estimation, safety, and fleet management purposes. For example, projects investigated equipment detection [2], [4], [5], [6], tracking [3], [7], [8], [9], activity interpretation [3], [7], [10], and safety assessment [17], [18]. These investigations mentioned some advantages to using vision-based monitoring systems, including low cost and the non-intrusive nature of this monitoring approach. Despite all the achievements, a number of shortcomings exist in the developed vision-based methods that prevents their practical applications. These limitations include those described below.

Low level of reliability: the rates of false positives and false negatives are still high compared to the radio-based sensing systems.

Line of sight and occlusion: vision-based systems require clear sightlines for the successful detection and tracking of operations, and obstructions can hinder the performance of the system. However, methods have been developed to improve the detection and tracking of semi-obscured targets.

Limited coverage: the developed systems use videos captured by stationary cameras. This limits the coverage area and will be problematic for large construction and open-pit mining fields. In those types of settings, site personnel should be able to set the camera’s view as the work progresses.

This research paper presents some methods for automated steering of a pan-tilt-zoom (PTZ) camera to overcome the coverage limitations of stationary cameras. In particular, it introduces intelligent steering frameworks for three main purposes: First, it demonstrates an algorithm to automatically scan the site by steering the camera and then finding active operations. Second, an algorithm is introduced to continue tracking an object when it is about to exit from the static view. Third, it discusses methods for identification and localization of equipment. All of these methods use image and video processing techniques to extract spatiotemporal data from 2D video frames, which are then used to steer the PTZ unit. Finally, opportunities and future research directions are discussed.

2 Research Methods

This section introduces methods developed to steer a PTZ camera for various purposes, such as resource tracking and automated operation localization in jobsites. These frameworks employ combinations of computer vision algorithms, such as image stitching, object recognition and feature tracking; active control of a PTZ camera; and principles of camera geometry to achieve the expected performance. The following subsections explain the technical approaches used in these systems.

2.1 Automated Search for Active Operations

Stationary cameras cover a limited area, which is especially problematic for large construction jobsites. For full coverage, the viewfinder of the camera must be adjusted manually as the work progresses. However, this issue could be tackled using a self-steering camera that searches the jobsite for an active operation. The first step of this process is to determine the boundaries of the camera's coverage area. In this step, the position of the camera, including its 3D coordinates and the initial yaw and pitch angles of the camera mount, is mapped to the jobsite plan. Then the boundaries of the site are considered to avoid scanning areas outside of the site boundaries (see Figure 1). In this arrangement, the optical axis is not allowed pass the boundaries of the jobsite.

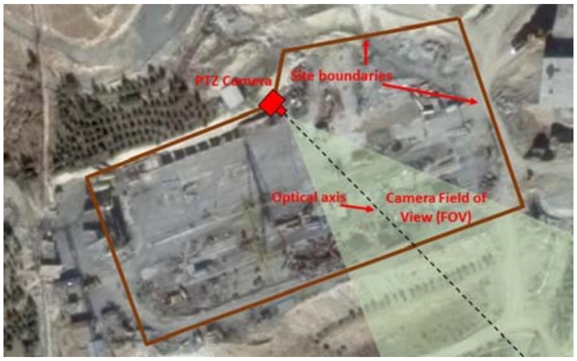

Figure 1. Mapping the camera in the site plan

The next step is the systematic search within the assigned boundary for active operations. An image stitching technique was used to create a panorama of the jobsite from several overlapping images. In this approach, the system changes the yaw angle of the mounting unit in a certain step and captures a frame. Then it steers the camera to the next viewpoint, in which it has an overlapping field of view (FOV) with the previous frame.

After the required frames are captured, invariant local features are used to find matches between frames and stitch those together [11]. This algorithm finds and matches SIFT features [12] in frames, and then employs a probabilistic method to verify the matches. Afterwards, it solves and refines camera parameters, compensates for errors, and blends the images (Figure 2. a).

The next step in this process is to search the panorama for construction equipment (Figure 2 b). Methods for recognizing rigid [4] and articulated equipment (such as excavators) [5] have been discussed. If detectors locate potential targets in the panorama, the system steers the PTZ camera to the located targets to verify detection. Sometimes the detectors provide false positives, so spatiotemporal information is used to help eliminate them. Some consecutive frames are searched for the target, then the spatiotemporal data of the detections are compared with the movement pattern of active equipment to reject false positives [5] [13].

When the detectors find different types of equipment, stationary plants have priority over highly mobile equipment in identifying jobsite tasks. For example, an excavator is an indicator for an excavation/loading operation, whereas a dump truck could be seen in different parts of a construction site.

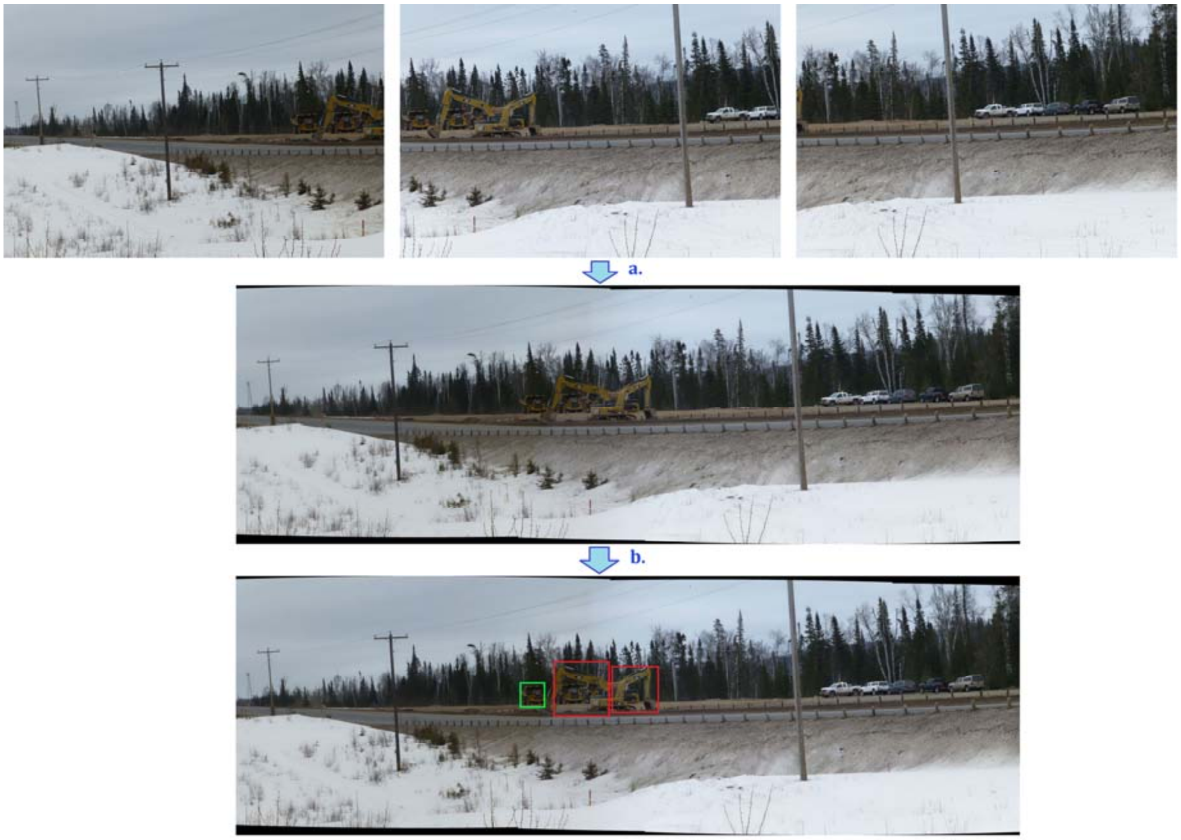

Figure 2. Steering of a PTZ camera to locate operations: a) stitching of images to create a panorama, b) object detection

2.2 Active Camera Control to Track Equipment

The tracking performance of the existing vision-based systems is limited to the camera's FOV. But some applications, such as security and safety monitoring systems, might need to track a target beyond the FOV of a stationary camera. Thus, it is useful to benefit from the physical panning and tilting capabilities of a PTZ camera to extend its tracking range.

A key issue for continuous tracking of construction resources is to estimate their motion vector, which includes velocity and trajectory. The velocity of the object is measured in the frame coordinate system, thus it would be expressed in terms of pixels per second. Then the motion vector of the target is used to predict the upcoming 2D coordinate of the target. The trajectory and velocity of the detected objects could be estimated using a tracking algorithm. The KLT feature tracker [14] was utilized to track certain features of the target object. The KLT is a differential method that uses spatial intensity data to search for the best match in the next frame. Successful motion estimation requires a number of specific interest points that are mathematically well-founded. Corners (the intersection of the edges) are suitable visual features for tracking, and Harris corner detection [15] was used to detect a number of corners for tracking (see Figure 3.b). Tracking each of these features yields a motion vector (see Figure 3.c). The overall motion vector of the target is calculated as the average of the valid motion vectors. The random sample consensus method (RANSAC) was used to remove the outlier vectors, vectors that indicate a direction inconsistent with that of most of the other optical vectors. Since construction equipment provides slow-moving targets in videos, the system estimates the upcoming coordinates of a piece of machinery using its current coordinates in the frame and its speed and trajectory. Then the system should steer the camera mount to maintain the target in the center of the frame. Given the pixel distance between current and predicted positions of the target, it is possible to determine also the pan and tilt angles to reach the projected position. This could be calculated using the horizontal and vertical angles of camera views. These angles are determined based on the size of the sensor and focal length of the camera's lens as presented in Equation (1):

$$\alpha = 2arctan\frac{d}{2f} \quad (1)$$

α is the angle of view (horizontal, vertical, or diagonal)

d is the length of the sensor in the direction measured (horizontal, vertical, or diagonal)

f is the focal length

Given the horizontal and vertical angles of camera views, the required displacement of the frame's center, and frame dimensions in pixels, the camera steering pan and tilt angles are calculated and executed (see Figure 3.d).

This algorithm, however, could consider the speed of the camera's panning and tilting system to achieve the best results. Most camera mounting systems are quite fast (more than 180°/Sec in pan and 60°/Sec in tilt movement) and don't cause any latency.

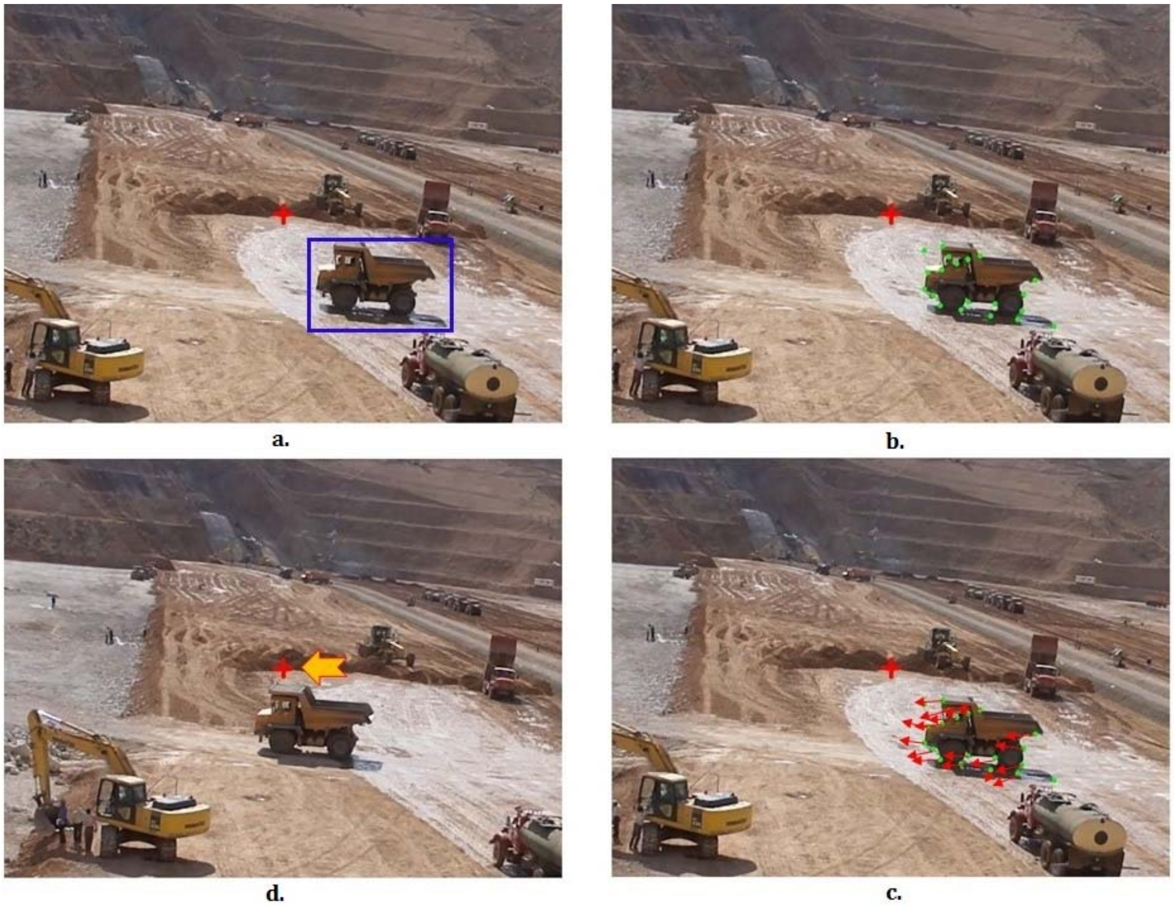

Figure 3. Camera steering process: a) detection of the target, b) generation of suitable features to track, c) optical flow estimation, d) panning of the camera to a new position

2.3 Zooming for Equipment Identification

Sometimes the zoom of the camera needs to be changed to obtain a suitable resolution of the targets. For example, the visual markers attached to pieces of equipment for identification and localization purposes should have a specific minimum size in the frame (in pixels) to be robustly detected and provide useful localization results. On the other hand, sometimes it is necessary to have a wide-angle view of all of the processes for tracking and monitoring purposes.

One of the main shortcomings of the model-based object recognition methods is that they cannot differentiate individual pieces of equipment within a same class. For example, they are not able to identify individual dump trucks within a fleet of similar machines. Identification of individual equipment is an important parameter for productivity estimation. The solution for this problem is to attach unique visual markers to every piece of equipment.

For detection of visual markers, such as AprilTag [19], attached on the equipment, the size of the marker in the frame (in pixels) is the key factor. Therefore, an automated zooming feature was developed to control the focal length of the lens. In particular, this module was developed to reliably detect markers attached to the equipment. Experiments revealed that an AprilTag marker should be at least 18x18 pixels to be reliably detected (>98%) [16].

An automated zoom control algorithm was developed to change the focal length of the camera's lens to obtain the expected marker resolution. First, the linear relationship between the focal length and the magnification factor of the camera lens was determined using a set of experiments. For this purpose, a specific target was captured in frames from the same viewpoint but with altered focal lengths. Then the linear relationship was established by measuring the changes of the target size (in pixels) versus the focal length. Second, the algorithm calculates the magnification factor required to obtain expected marker size (in pixels) based on the size of the detected equipment (in pixels), and sets the focal length of the lens to achieve the expected resolution. This algorithm calculates the magnification factor using the ratio of the actual length of the visual tag to the actual length of the equipment (e.g. 0.6m label and 7m dump truck length) and the expected marker size (e.g. 20x20 pixels). For example, a dump truck is detected with a size of 190x119 pixels in a frame, thus the size of the marker is about 16x16 pixels in the frame (190x(0.6/7) = 16). Assuming that the expected marker resolution is 20x20 pixels, the required magnification factor would be 20/16=1.25.

In addition, to avoid useless zooming, the algorithm checks whether the target will remain in the view after zooming. This test is done using a scale affine transformation, in which the coordinates of the object after magnification are calculated. The new coordinates indicate whether the target will appear within the zoomed frame.

Since the camera lens zooms toward the center of the view, the center of the frame is set as the origin of the 2D coordinate system. Then the coordinates of the detected objects are multiplied by the calculated magnification factor to check whether the target will appear in the frame after zooming. An autofocus process (usually provided by the camera manufacturer) is employed to correct the usual blurriness resulting from change of the focal length.

The zooming module was used to identify dump trucks and excavators that were labelled with unique fiducial markers, and the reported results are presented in Table 1. Due to the robust performance of the marker recognition algorithm, no false positives were observed [16].

Table 1. Performance of the equipment identification system using fiducial markers and active zooming

Experiment	No. appeared machines	No. identified machines	Precision	Identification rate
Dump truck	32	27	100%	84.4%
Excavator	18	16	100%	88.9%

2.4 Active Camera Control to Localize Equipment

Automated camera steering could be used to estimate 3D location of the targets. In this approach, the PTZ camera is the origin of a spherical coordinate system. The three parameters—the radial distance of the target from the camera, its tilt (polar) angle measured from the zenith direction, and pan (azimuthal) angle measured by its projection on the horizon plane from a fixed reference direction—are required to estimate the 3D coordinates of an object with respect to the PTZ camera. The pan and tilt angles are provided by the motorized camera mounting systems, thus the only parameter requiring estimation methods is "the radial distance"—the distance of the target to the camera.

Depth estimation using a single camera could be resolved using two main approaches: one based on the geometric principles of a pinhole camera and one based on a pinhole camera model. These methods also rely on detection of a fiducial marker attached to the equipment. So the system needs to zoom on the target for localization of the target.

The first step in both approaches is to calibrate the camera to find the focal length(s) and principal point of the camera. Then, the vision system searches the video frames for the objects of interest, i.e. construction equipment. The object recognition methods were discussed in previous works [4], [5]. As soon as a target is spotted in the frame, the system steers the camera to capture the target in the center of the frame to achieve the best depth estimation result. The process of the camera steering to capture the target in the center of the frame is similar to the method described in section. 2.2. Then the system zooms (by changing the focal length) on the target to detect the marker attached on the equipment. The following subsections describe the two depth estimation approaches.

2.4.1 Camera Geometry

The first technique is to use the geometric principle of a pinhole camera. In Equation (2):

$$d = \frac{f \times rho \times imh}{rhi \times seh} \tag{2}$$

d is distance to the camera in mm
f is the focal length in mm
rho is the real height of the object in mm
imh is the image height in pixels
rhi is the object height in image in pixels
seh is the camera's sensor height in mm

This method achieves the best result when the target is parallel to the image plane, and it is less successful when the planar target is rotated with respect to the image plane.

2.4.2 Pinhole Camera Model

This approach is based on the pinhole camera model, in which the intrinsic and extrinsic parameters of the camera are used to estimate the distance from a planar target to the camera (see Equation (3)). The intrinsic matrix includes the camera's focal length, principal point, and distortion parameters, which is extracted using a well-known camera calibration process.

$$sm = A[R|T]M \tag{3}$$

s is a scale factor. This coordinate system is a homogenous system and scale invariant.

m is a 3x1 matrix of the pixel coordinates of the projected point in the image. The third element (z) is equal to 1.

A is the camera's intrinsic matrix.

[R|T] is the 3x4 extrinsic matrix. R is a 3x3 rotation matrix demonstrating the orientation of the camera with respect to the planar target, and T is a 3x1 translation matrix which denotes the camera's position.

M is a 4x1 matrix which includes the 3D world coordinates of the point. The fourth element is equal to 1. To simplify the calculations, a point on the planar object is usually set as the origin of the world coordinate.

This approach handles rotation of the targets theoretically, by considering the rotation of the planar target. In this method, a corner of the target (i.e., the visual marker) will be the origin of the world coordinate and the target plane is set as the world's x-y plane; given the dimensions of the target, the world coordinates of the rest of three corners are calculated. Then four equations (based on Equation (3)) are established for the four corners, in which all parts except [R|T] matrix, including m, A, and M, are known. Using a set of linear equations, the [R|T] matrix is extracted [20].

Afterwards, the [R|T] matrix is used in Equation (3) to estimate the 3D coordinates of the camera origin in the 3D coordinate system (where its origin is on the target). Lastly, the norm of the camera's 3D coordinate yields the distance between the target and camera.

2.4.3 Experimental Results

This localization framework was evaluated using a number of real-time videos of different earthmoving equipment, including excavators and off-highway and urban dump trucks. A total station was set up adjacent to the PTZ camera for accurate measurement of the location of the targets. These videos included 67 test samples, in which the distance of the test subjects to the camera varied from 30.44 to 100.50 meters. A few samples are presented in Figure 4. Table 2 provides the performance of localization methods. As it is presented, the results are relatively close, but the pinhole camera model performed slightly better. Comparison of the results demonstrated that the performance of the pinhole camera model approach was better than the camera geometry method in localization of the targets that were oriented with respect to the image plane.

Table 2. Performance of the localization algorithms

Localization method	Average localization error (m)	Standard deviation of error (m)	Maximum error (m)	Upper limit at 95% confidence level (m)
Camera geometry	1.389	0.992	4.18	1.631
Pinhole camera model	1.346	1.031	4.11	1.597

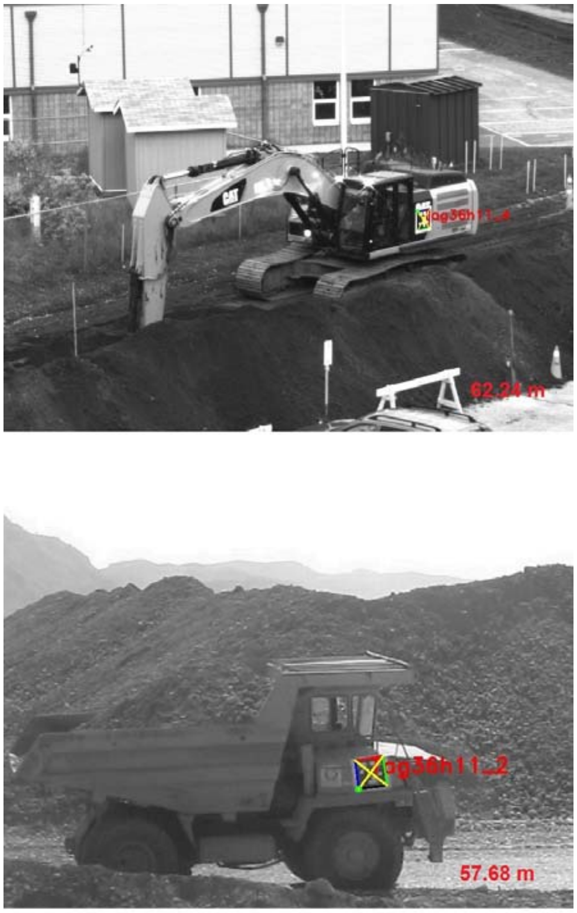

Figure 4. Samples of marker localization

2.5 Conclusion

Vision-based systems are among the sensing technologies that are increasingly being tested to monitor the productivity and safety of equipment-intensive operations in the construction and surface mining industries. The developed systems, however, analyze videos captured by a motionless camera, which drastically limits their coverage area. This research paper introduces some algorithms to automatically control pan, tilt, and zoom features of a PTZ camera. These camera steering algorithms command the PTZ unit based on the spatiotemporal data extracted from the video frames. The methods discussed include automated scanning of the site for active operations, continuous tracking, and identification and localization of the labelled equipment.

The discussed methods, however, are able to control only a single camera, which is not sufficient for large construction and mining fields. Future research will investigate autonomous control of a network of connected PTZ cameras to monitor large jobsites.

References

[1] Yang J., Park M.W., Vela P.A. and Golparvar-Fard M. Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future. *Advanced Engineering Informatics*, 29(2):211-224, 2015.

[2] Memarzadeh M., Golparvar-Fard M. and Niebles J. C. Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. *Automation in Construction*, 32:24–37, 2013.

[3] Gong J. and Caldas C.H. An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations. *Automation in Construction*, 20(8):1211–1226, 2011.

[4] Rezazadeh Azar E., and McCabe B. Automated visual recognition of dump trucks in construction videos. *Journal of Computing in Civil Engineering*, 26(6):769-781, 2012.

[5] Rezazadeh Azar E. and McCabe B. Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos. *Automation in Construction*, 24:194-202, 2012.

[6] Chi S. and Caldas C.H. Automated Object Identification Using Optical Video Cameras on Construction Sites. *Journal of Computer-Aided Civil and Infrastructure Engineering*, 26(5):368–380, 2011.

[7] Rezazadeh Azar E., Dickinson S. and McCabe B. Server-Customer Interaction Tracker: Computer Vision–Based System to Estimate Dirt-Loading Cycles. *Journal of Construction Engineering and Management*, 139(7):785–794, 2013.

[8] Park M., Koch C. and Brilakis I. Three-Dimensional Tracking of Construction Resources Using an On-Site Camera System. *Journal of Computing in Civil Engineering*, 26(4):541–549, 2012.

[9] Park M.W., Makhmalbaf A. and Brilakis I. Comparative study of vision tracking methods for tracking of construction site resources. *Automation in Construction*, 20(7):905-915, 2011.

[10] Golparvar-Fard M., Heydarian A. and, Niebles J.C. Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers. *Advanced Engineering Informatics*, 27(4):652–663, 2013.

[11] Brown M. and Lowe D. G. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59-73, 2007.

[12] Lowe D. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] Rezazadeh Azar E. and McCabe B. A visual sensing approach to estimate material hauling cycles in heavy construction and surface mining jobsites. In *Proceedings of the ISARC 2013 conference*, paper 229, Montreal, Canada, 2013.

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

[15] Harris, C. and Stephens, M. A combined corner and edge detector. In *Proceedings of the Alvey vision conference*, 15, page 50, 1988.

[16] Rezazadeh Azar E. Construction Equipment Identification Using Marker-Based Recognition and an Active Zoom Camera. *Journal of Computing in Civil Engineering*, 04015033, 2015.

[17] Chi S. and Caldas C. H. Image-based safety assessment: Automated spatial safety risk identification of earthmoving and surface mining activities. *Journal of Construction Engineering and Management*, 138(3):341–351, 2012.

[18] Kim H., Kim K. and Kim H. Vision-Based Object-Centric Safety Assessment Using Fuzzy Inference: Monitoring Struck-By Accidents with Moving Objects. *Journal of Computing in Civil Engineering*, 04015075, 2015.

[19] Olson E. AprilTag: a robust and flexible visual fiducial system. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA'11)*, pages 3400–3407, Shanghai, China, 2011.

[20] Zhang Z. A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11):1330-1334, 2000.