Automated visual recognition of rigid earthmoving equipment in construction videos

Ehsan Rezazadeh Azar¹ and Brenda McCabe²

Abstract

Earthmoving plants are essential but costly resources in the construction of heavy civil engineering projects. In addition to proper allocation, ongoing control of this equipment is necessary to ensure and increase the productivity of earthmoving operations. Captured videos from construction sites are potential tools to control earthmoving operations; however, the current practice of manual data extraction from surveillance videos is tedious, costly and error prone. Cutting edge computer vision techniques have the potential to automate equipment monitoring tasks. This paper presents research in the evaluation of combinations of existing object recognition and background subtraction algorithms to recognize earthmoving machines in video streams. Two detection algorithms, namely Haar-HOG and Blob-HOG, are presented and evaluated for their ability to recognize off-highway dump trucks in videos as measured by both effectiveness and timeliness. The results of this study can help practitioners select a suitable approach to recognize construction equipment in videos for real-time applications such as productivity measurement, performance control, and proactive work-zone safety.

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

Introduction

Earthwork projects are a major segment of the construction industry and they depend on large fleets of heavy equipment. The operations involved in earthmoving projects, including open pit mining, are very repetitive; slight improvements in cycle times can result in significant improvements in productivity and cost savings. Therefore, a number of studies have been carried out for both planning and controlling these operations. Simulation (Cheng et al. 2011; Kim and Kim 2010), Bayesian belief networks (McCabe and AbouRizk

¹ Ph.D. Candidate, Department of Civil Engineering, University of Toronto, 35 St. George St., Toronto, ON. M5S1A4, Canada, rezazade@ecf.utoronto.ca
² Associate Professor, Chair, Department of Civil Engineering, University of Toronto, 35 St. George St., Toronto, ON. M5S1A4, Canada, brenda.mccabe@utoronto.ca
2001), genetic algorithms (Marzouk and Moselhi 2004), neural networks (Shi 1999), and linear programming (Moselhi and Alshibani 2009; Son et al. 2005) have been applied to optimize equipment allocation or to find the bottlenecks of these operations in advance.

On the other hand, controlling site operations in real-time requires person to person communication with equipment operators/supervisors and/or watching the operations directly which is costly and error-prone. The hearing sense can be mimicked and automated using technologies that capture radio/sound signals such as global positioning system (GPS) satellite receivers, ultra wideband (UWB) systems, equipment sensors, or radio frequency identification (RFID) tags. The earthmoving sector currently uses many of these technologies to control their operations, which has made this sector the frontier of automated performance control in construction (Navon 2005). Several research projects have been carried out to evaluate and improve the efficiency of these tools. For example, GPS can send the geographical location of machines to a control unit at regular time intervals. These location data together with logic algorithms of the control system can recognize the type of activity and measure the productivity of the machine (Navon et al. 2004; Kim and Russell 2003). One disadvantage of these systems is that each piece of equipment to be tracked will require the appropriate technology to be installed and kept up to date. If contractors own, rent or purchase equipment that does not already have the appropriate tracking technology, then it must be installed.

Computer vision-based systems are an alternative for visual monitoring and tracking of earthmoving equipment. Due to the emergence of high performance processors, there has been a dramatic increase in applications for vision-based systems in different sectors of the industry. Construction researchers have also studied the application of image processing techniques for several construction engineering tasks including building progress monitoring (Wu et al. 2010; Golparvar-Fard et al. 2009), productivity measurement (Gong and Caldas 2011; Weerasinghe and Ruwanpura 2009), defect detection (Zhu and Brilakis 2010; Guo et al. 2009), and automated documentation (Brilakis and Soibelman 2005; Brilakis et al. 2005). Computer vision techniques are also adaptable in earthwork monitoring systems, because earthmoving machines usually operate in relatively visible areas and, unlike the other types of construction projects such as building and industrial projects, the scenes are often not obstructed. As a result, some researchers investigated the potential of computer vision algorithms to monitor earthmoving plants; for example, Makhmalbaf et al. (2010) evaluated the performance of object tracking methods for equipment and resources in construction videos. This research studied tracking algorithms where the targets were manually selected in the frames and passed to tracking software. In another effort, image color space was used to detect a hydraulic excavator and measure its idle time in time-lapse images with earth material and snow as a background (Zou and Kim 2007). A background subtraction method was applied to detect moving objects in construction videos with static background, and then these detected objects were identified by Bayes or neural network classifiers (Chi and Caldas 2011a). This recognition system was then applied for safety risk identification of earthmoving activities (Chi and Caldas 2011b).

Since heavy equipment is the main operational resource in earthmoving projects, vision-based algorithms target those machines to control the operations. Vision-based monitoring
systems consist of detection and tracking algorithms to provide 2D location of the machine at any point in time. It is further possible to calculate the 3D coordinates of any point by using more than one camera view and an epipolar geometry algorithm (Makhmalbaf et al. 2010; Zhu and Brilakis 2009).

Detection of equipment in videos is beneficial in several ways. First, it becomes possible to count the number of trucks passing the camera view to estimate dirt volume and measure the time frame of each service for productivity measurement. Second, this system can locate the idle or waiting machines in queues and notify the site engineer of the situation in real time. Other applications, such as accident warning and intelligent image and video indexing, can benefit from automated recognition system. Low cost and shorter preparation times are the main advantages of the vision-based systems over the hearing-based systems. Further, vision systems are less intrusive in that they do not require the same technology to be installed on every item to be tracked.

This paper presents research into an automated vision-based monitoring system to detect rigid-frame equipment in digital videos. Two cascade algorithms to detect equipment in construction videos are presented and compared. Along with detection precision, run-time efficiency for real-time application is considered equally important. The scope of this paper is limited to object recognition and tracking methods are not covered here. The main contribution of this paper is a rapid and accurate equipment recognition framework for construction videos.

**Object Recognition**

Object detection and recognition, which deals with finding objects of interest in digital images and videos, is one of the main streams in computer vision research. Vision-based object recognition has attracted much interest and several algorithms have been developed in the last two decades. For this research, several existing algorithms were investigated and compared for their ability to detect objects similar in shape to dump trucks, such as urban vehicles. The evaluation criteria were detection rate, false positives, and runtime efficiency, i.e. computation speed. See Table 1 for the possible classification outcomes. There are different rules about the acceptance of a true positive, but for this research, the evaluation rules of the PASCAL visual object classes challenge (Everingham and Winn 2010) were followed, which requires the detected bounding box to overlap more than 50% with a ground-truth bounding box to be considered a true positive. Other detected boxes or multiple detections are counted as false positives. Two algorithms, namely HOG and Haar, were selected and are introduced in next sections.

**Histogram of Oriented Gradients (HOG)**

One of the cutting edge algorithms, Histogram of Oriented Gradients (HOG) (Dalal and Triggs 2005), won the 2006 PASCAL object detection challenge (Everingham et al. 2006). Because of its robust performance in the detection of rigid objects, this method has been used for automated vehicle recognition and traffic safety (Rybski et al. 2010; Morlock 2008). It was also used in the development of part based models (Felzenszwalb et al. 2010), which won the 2009 PASCAL object detection challenge (Everingham et al. 2009). The HOG detector uses the SVMLight algorithm (Joachims 1999) to create a vector of features
from a training dataset as it searches for the target object in the test images. To create a vector, computed gradients of the gray-scaled image are discretized into spatial and orientation bins. Figure 1 shows a sample image and graphical representation of its HOG feature vector computed from a downscaled 128x80 pixel image of the original photo. In the visualized HOG image in the right hand side, the gray-scaled frame is divided to 4x4 rectangular blocks, and then the computed gradients of each block are grouped into nine orientation bins that are evenly spaced over 0°-360° (unsigned gradient). This method highlights the external and internal edges of the dump truck.

Table 1. Description of Four possible Outcomes of a Classifier

<table>
<thead>
<tr>
<th>Outcome of classification</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive</td>
<td>Object is correctly recognized</td>
</tr>
<tr>
<td>False positive (false alarm) or Type 1 error</td>
<td>Wrong window misclassified as the target</td>
</tr>
<tr>
<td>True negative</td>
<td>Nonobject window correctly rejected</td>
</tr>
<tr>
<td>False negative or Type 2 error</td>
<td>Object is missed</td>
</tr>
</tbody>
</table>

Figure 1. Left: Original image, Right: Visualization of the HOG descriptor

**Haar-like features**

A very rapid and popular object recognition framework, Haar-like features (Viola and Jones 2001), was also considered. The Haar-like features method is a cascade detector that employs a series of weak classifiers in which each of the classifiers applies dense encoding of image regions similar to Haar wavelets basis functions. Any test window that cannot pass one of the classifiers is not sent to the next weak classifiers; therefore, unlike the HOG method it does not completely process all the search windows with all of the classifiers which results in much faster processing times. This object detection algorithm uses a form of the AdaBoost learning algorithm (Freund and Schapire 1997) to both select the best features and train the classifiers.

**Data Preparation**

**Data collection**

The first step of this research was to collect images and videos from the multimedia archives of a heavy-civil construction company and freely-available online resources. In
addition, images and videos were captured by one of the authors from a large rock-fill dam construction project. All of the photos and videos were taken in daylight with wide-ranging levels of illumination. Images and videos were captured from ground and above ground angles and the equipment in them were located at various distances from the camera. Many of these images contained different makes and sizes of dump trucks (see Table 2 for more details) and were used either for training or testing purposes. Also, some truck-free images were collected and used as a negative data set. Videos were only used to assess the performance of the detectors (i.e. not for training).

Table 2. Types and sizes of dump trucks appeared in the collected images

<table>
<thead>
<tr>
<th>Brand</th>
<th>Model</th>
<th>Payload (tonnes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caterpillar</td>
<td>777D</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>775E</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>773B,D</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>769C,D</td>
<td>36</td>
</tr>
<tr>
<td>Komatsu</td>
<td>HD325</td>
<td>36.5</td>
</tr>
<tr>
<td></td>
<td>HD465</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>HD605</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>HD785</td>
<td>91</td>
</tr>
<tr>
<td>Belaz</td>
<td>7540</td>
<td>30</td>
</tr>
<tr>
<td>Euclid</td>
<td>R65</td>
<td>61.5</td>
</tr>
<tr>
<td></td>
<td>R50</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>R35</td>
<td>35</td>
</tr>
<tr>
<td>Terex</td>
<td>TR100</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>TR60</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>TR45</td>
<td>41</td>
</tr>
</tbody>
</table>

Classifying visual orientations

Dump trucks, like other vehicles, appear quite different depending on the viewpoint with respect to the camera. This makes anything but a sphere impossible to be recognized with a single classifier. Previous research showed that the application of eight viewpoints had robust results in vehicle recognition (Rybski et al. 2010, Han et al. 2006), thus the same approach was followed and the training dataset was grouped into eight orientations as illustrated in Figure 2. The application of eight orientations also enabled identification of the trajectory of the dump truck, which is useful for accident warning and activity recognition systems.

Training the detectors

A large number of dump truck images were randomly picked from the image dataset for training of the detectors. These training images were used to train both Haar and HOG classifiers. Table 3 displays the number of training images in each of the eight orientations.
All of the positive and negative training samples were manually cropped and then scaled to predetermined sizes. The Haar and HOG methods use different training approaches, thus the settings are different with the Haar method requiring relatively small sizes such as 20x20 to 40x40 pixels, but the HOG algorithm requires a bigger search window such as 64x128 or 104x80. Therefore, different training window sizes were used for each method as presented in Table 4. The main issue is to set a proportion for the bounding boxes to completely enclose the machines, but addition of 16 pixels of margin around the target object on all four sides showed the best performance for HOG detectors (Dalal and Triggs 2005). We considered 72x64 for front and rear views and 96x48 windows for the rest orientations to enclose the positive truck samples, and after addition of 16 pixels of margin in each side, the results would be 104x96 and 128x80 boxes. However, the positive training samples for Haar method need smaller windows with smaller margins. Thus, firstly 10 pixels were cropped from each side of the HOG’s positive samples to reduce the margin, and the remaining frames have the sizes of 84x76 for front and rear views and 108x60 for the rest of orientations. Afterwards, 84x76 boxes were resized by 1/4 scale factor and 108x60 images by 1/3 which resulted in 21x19 and 36x20 windows respectively.

Many of the 800 negative images contained not only blank construction landscapes but also other earthwork machines, such as excavators and bulldozers, to help reduce misclassification of such equipment. Ten windows were randomly cropped from each image and scaled to corresponding category sizes, which resulted in 8000 negative training samples for each viewpoint. The HOG method requires two rounds of training for better
performance; the firstly trained classifier searched the unscaled 800 negative images and all of the detected windows (which were definitely false positive) were scaled and added to the initial 8000 negative images for a second round of training. These few additional hard negatives can improve the results of the final classifier (Dalal and Triggs 2005). Open source OpenCV 2.1 (OpenCV 2010) library was used to train Haar classifiers and calculate HOG descriptors, and publicly available SVMlight (Joachims 1999) software was used for training the HOG classifiers.

**Experiment on Static Images**

**HOG detectors**

The detection process includes scanning of each test image with eight single-class (one for each viewpoint) classifiers. To evaluate the performance of the detectors, 380 test images were randomly selected from the most challenging images, none of which were used for training. In all, 681 dump trucks appeared in all eight orientations together with other types of heavy equipment, some of them with similar colors, to assess the performance of the classifiers in congested views. The receiver operating characteristic (ROC) curve of this detection experiment is illustrated in Figure 3. The ROC curve graphically shows the detection rate versus the false positives per frame at varying thresholds. The detection rate is defined as the proportion of actual equipment correctly identified, or (true positives)/(true positives + false negatives). As the HOG classification threshold decreases, more test windows, including true and false positives, will pass the SVMlight classification process. A detected object is only accepted as a true positive if the classifiers correctly determine both the location and the orientation of the dump truck in the image. For instance, if the detector locates a dump truck with “Front-right” orientation instead of “Rear-right”, it will be considered a false alarm.

Some recognition samples are presented in Figure 4. As observed in the outcomes, this method could robustly distinguish the dump truck from other types of equipment, of which many have a similar color. For example, Figure 4 shows images taken from a rock-fill dam construction project with bulldozers, rollers, loaders, graders, and hydraulic excavators. A roller is misclassified in “side-right” orientation in the top right image.

![ROC curve](image1.png)

Figure 2. ROC curve of the HOG detector
Table 4. Training details of each method

<table>
<thead>
<tr>
<th>Method</th>
<th>Front and rear views (pixels)</th>
<th>Other six views (pixels)</th>
<th>Training algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haar-like</td>
<td>21x19</td>
<td>36x20</td>
<td>SVMlight</td>
</tr>
<tr>
<td>HOG</td>
<td>104x96</td>
<td>128x80</td>
<td>Gentle AdaBoost</td>
</tr>
</tbody>
</table>

Because the test samples were randomly selected from the most challenging images, several dump trucks were partially masked with soil depots or other machines which caused most of the missed objects (see Figure 5). On the other hand, many of the false positives resulted from wrong orientation estimations rather than wrong location estimations.
In addition to detection rates and false alarms, the computation times for different sizes of images are presented in Table 5 to assess the possibility of real-time applications. These runtimes are obtained on a dual core 2.93 GHz CPU. Scanning all eight viewpoints in a low resolution standard surveillance image of 640x480 pixels takes about 26 seconds, which is too long for real time processing. The reason is that classifier first searches for the target in the original scale image, then the image is scaled down by the shrinkage coefficient (set at 1.05 in this experiment), and the scan is repeated. This process continues until the image reaches the size of the classifier, which were 128x80 and 104x96 pixels (Table 4). For example, to find dump trucks in all eight orientations in a 640x480 image, the system should classify 6x40508 or 243,048 windows for the six orientations with the 128x80 classifier and 2x40999 or 81998 windows for “Front” and “Rear” viewpoints, which have 104x96 search windows. Each of these SVMlight classification processes is a scalar product of the classifier and the test window vectors for which the vector sizes are [4752x1] for 104x96 windows and [4860x1] for 128x80 boxes.

Scaling down the frames before running the detection process is one option to reduce the runtime; however, the risk of missing objects in the images increases dramatically, because any dump truck smaller than the search window cannot be recognized by the detectors.

Table 5. Run time of the HOG detectors in scanning of eight orientations

<table>
<thead>
<tr>
<th>Image size</th>
<th>HOG (Sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>640x480</td>
<td>26</td>
</tr>
<tr>
<td>1024x768</td>
<td>69</td>
</tr>
<tr>
<td>1920x1080</td>
<td>186</td>
</tr>
<tr>
<td>2592x1944</td>
<td>455</td>
</tr>
</tbody>
</table>

*Haar-like detectors*

The performance of the standalone Haar method was disappointing in the detection of dump trucks. Despite its fast performance (Table 6), it showed relatively low detection and
high false positive rates compared to the HOG technique (see Table 7). As such, it was temporarily set aside as a standalone detector.

Table 6. Run time of the Haar detectors in scanning of eight orientations

<table>
<thead>
<tr>
<th>Image size</th>
<th>Haar (Sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>640x480</td>
<td>1.2-2.0</td>
</tr>
<tr>
<td>1024x768</td>
<td>2.5-5.1</td>
</tr>
<tr>
<td>1920x1080</td>
<td>6.9-13.1</td>
</tr>
<tr>
<td>2592x1944</td>
<td>19.6-25.7</td>
</tr>
</tbody>
</table>

**Experiment on Videos**

Object detection in videos is a more challenging task than in static images because most video applications are real-time, such as operation monitoring and safety controlling. Therefore, the algorithms should be rapid enough to be processed on standard computers. As shown in Table 5, the HOG run-times on even relatively low resolution frames are too long for real-time applications. One of the common strategies in computer vision field to speed up the classifications is cascade approach, which first applies a rapid and probably weak filter to eliminate object-free regions and pass only the most probable candidates to a robust and slower classifier (such as HOG) for final verification. Training flexible HOG block sizes by the Adaboost algorithm to create a rejection cascade is an instance of the cascade approach for fast pedestrian detection (Zhu et al. 2006). In a system designed to detect people in static images, the combination of Adaboost and HOG was used to filter ‘head and shoulder’ candidates, which were sent to another Adaboost classifier to confirm the entire body (Zhou et al. 2009).

Motion detectors can capture moving objects in a video stream and send them to a robust detector for final classification. For instance, optical flow was used as a motion descriptor and HOG for the appearance descriptor to detect pedestrians in videos (Dalal et al. 2006). The next section describes the performance of two algorithm combinations developed here to speed up the recognition process on test videos.

**Haar-HOG cascade method**

The Haar-HOG method consists of a Haar detector as a fast classifier to eliminate most true negatives and reduce the number of search windows; then a HOG detector is used to reject the false alarms and keep true positives. At first glance this combination may look similar to a cascade algorithm which was developed to detect the cars ahead of the driver (Negri et al. 2008), but the Haar and HOG combination was used to extract features and Adaboost technique was employed for learning and classification; however, in this research
each of the Haar and HOG detectors were trained and used separately. Figure 6 depicts the flowchart of this approach.

![Figure 6. Flowchart of the cascade algorithm](image)

For the Haar-HOG cascade method, the Haar detectors trained with higher false alarm rates including the ones with 0.55 and 0.6 thresholds for the maximum false positive rate were employed. The false alarm thresholds were set rather high, resulting in both high true positive and false positive rates (see Table 7). In contrast with HOG methods where the threshold is set in the classification stage, the detection thresholds of Haar are defined in the training stage, which determines the detection rate and max false alarm rate in processing the training samples. Training the Haar classifiers with various thresholds provided the opportunity to assess the detection rates and process times of the Haar-HOG method.

After training the Haar classifiers, test images were shown to this fast detector to produce a limited number of candidate windows (see Figure 7.a). Then, the HOG classifier checked these candidate windows using the same orientation to determine if it was really a truck (see Figure 7.b).

The challenge in the development of this algorithm was to synchronize the Haar and HOG methods. The HOG algorithm usually requires a bounding box that completely surrounds the object to detect it; but, there were several instances where the detection windows of the Haar method did not entirely enclose the trucks and as a result, the second classifier was unable to identify the truck. To solve this problem, the detection boxes of the Haar process were expanded to ensure that they fully enclosed the object. However, it was found that the expansion of the size of each box increased the search area, thereby increasing both the computation time and misclassifications. Thus, several expansion ratios were examined with a same set of classifiers and detection thresholds which Table 8 presents the outcomes. An expansion by 10% in all four sides of the bounding box had optimum results, improving
the detection rate by 16% with acceptable increase in runtime and false alarms. Although 15% expansion showed slight improvement, it increased runtimes and false positives significantly.

![Figure 7. a: Result of Haar detector; b: Result after checking of candidate windows by HOG](image)

The performance of the Haar-HOG algorithm on static images was evaluated on the same test dataset, and the results including ROC curve and runtimes are presented in Figure 8 and Table 9 respectively.

![Figure 8. ROC curves of HOG and cascade algorithms](image)

<table>
<thead>
<tr>
<th>Expansion ratio on each side (%)</th>
<th>Detection rate (%)</th>
<th>False alarms per frame</th>
<th>Average processing time of 640 x 480 pixel images (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>63.58</td>
<td>0.52</td>
<td>1.42</td>
</tr>
<tr>
<td>5</td>
<td>70.63</td>
<td>0.84</td>
<td>1.97</td>
</tr>
<tr>
<td>10</td>
<td>79.30</td>
<td>1.30</td>
<td>2.49</td>
</tr>
<tr>
<td>15</td>
<td>80.62</td>
<td>1.54</td>
<td>3.04</td>
</tr>
</tbody>
</table>
As shown in Table 9, the Haar-HOG method substantially accelerated the detection process compared to HOG alone; however, the detection rate was lower than HOG by about 8% (Figure 8). This weaker performance is due to the combination of the false-negative rates of both the Haar and HOG classifiers. The varied run-times of the Haar-HOG method on images of the same size were caused by different number of candidate windows generated by the Haar method in each photo. Even with this variation, a significant decrease in runtime was observed.

Table 9. Run times of HOG and Haar-HOG methods in scanning of eight viewpoints

<table>
<thead>
<tr>
<th>Image size</th>
<th>Detection runtime (Sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HOG</td>
</tr>
<tr>
<td>640x480</td>
<td>26</td>
</tr>
<tr>
<td>1024x768</td>
<td>69</td>
</tr>
<tr>
<td>1920x1080</td>
<td>186</td>
</tr>
<tr>
<td>2592x1944</td>
<td>455</td>
</tr>
</tbody>
</table>

To take advantage of the relatively rapid performance of the cascade method for real-time applications, the system should be set up to scan frame intervals slightly farther apart than the maximum runtime for that frame size. For instance, scanning each 640x480 pixel frame takes less than 2.6 seconds, thus the system can be set to examine video frames in ≥3 seconds intervals, so that the process can keep up with the real-time video stream. For this test, we set the algorithm to scan the frames every 4.7 seconds, because this time interval is a reasonable time period to detect a dump truck entering the scene and maintain run-time efficiency. As observed in the test videos, the dump trucks moved relatively slowly in the sites due to speed limits (25 km/h) and it took considerable time to pass the view of the camera, so all of the trucks were visible to the detectors in at least one frame.

The performance of this method was evaluated for the detection of off-highway dump trucks in test videos, all of which had 640x480 pixel resolution. It was generally agreed that frames do not need to have high resolution for heavy civil construction engineering applications and regular surveillance frame sizes such as 640x480 are sufficient to detect equipment with visual sizes bigger than the smallest classification windows (21x19 pixels for front and rear views and 36x20 pixels for the rest which then resized to corresponding HOG windows for final classification). Any dump truck smaller than these sizes in the frames will be missed; fortunately, most of the time they are not the vehicles of interest. All together, 17 videos with a total duration of sixty-five minutes were used to evaluate the performance of the cascade algorithm. These videos contained 64 dump trucks in different phases of their working cycles. One frame every 4.7 seconds was taken from the video stream and shown to the Haar-HOG detector, resulting in 828 frames being scanned. The Haar-HOG cascade detector results are shown in Figure 9 for 4 consecutive video test frames. The system processed all of the videos without any delay to the normal stream. Figure 10 illustrates the ROC of the results. In addition to ROC curve of Haar-HOG detectors, ROC curve of Blob-HOG method, which is described later in this paper, is plotted in the same graph for further comparison.
Figure 9. Object recognition in a series of frames at the specified time intervals (a through d)

Figure 10. ROC curves of both Haar-HOG and blob-HOG methods
Because the aim of this test was to detect trucks in the videos and not the performance of the detector in each frame separately, the results on the videos were evaluated differently which the detection rate is calculated as: number of detected dump trucks (regardless of frequency of detections)/number of appeared trucks in the video stream, therefore results are better than the static images. For example, if a dump truck appears in three frames and is detected in one, two or all three frames, it is counted as detected; however, any false detections in other frames were also counted. So the system had several chances to detect the 64 dump trucks. Due to the multiple opportunities for the detectors to identify the trucks, higher detection thresholds for second classifiers (HOG) were set than the ones used in static images, which brought about fewer false positives. In addition, the experiments on static images were carried out on different image sizes such as 1920x1080 and 2592x1944, which had many more search windows than the 640x480 pixel video frames; hence the probability of false detection was much higher in the static images than the videos.

**Blob-HOG cascade method**

The Haar-HOG algorithm finds objects in a static image, whether it is a single photograph or a frame from a digital video. In a second approach, the possibility of finding objects moving from frame to frame was investigated. In this method, moving objects are sent to an identifier for classification. To achieve this, a foreground-background segmentation filter is first applied to separate moving pixels from the static background. There are several background subtraction techniques available, but the selected algorithm should be able to acceptably process earthmoving videos with a continually changing environment and extensive visual noise such as dust and smoke from equipment exhaust. In a comparison of background subtraction algorithms (Gong and Caldas 2011), the performance of Codebook (Kim et al. 2005), Mixtures of Gaussian (Grimson et al. 1998), and a Bayesian-based model (Li et al. 2003) showed that the Bayesian model had better outcomes in construction videos; thus this algorithm was applied in this research. This foreground-background filter uses Bayes decision rules to detect both gradual and abrupt movements in videos captured by stationary cameras. However, the resultant image of this process is quite noisy and another process called connected-components analysis (Bradsik and Kaehler 2008) is required to remove random noise and integrate the correlated regions. This cleanup process applies a morphological operation to omit small noises that have smaller contours than a predetermined threshold, and retrieves the surviving areas and creates unified shapes called “blobs”. It then locates a bounding rectangle around each blob. The entire process of blob creation is shown in Figure 11.
Each of these blobs should represent the entire outline of a moving object; however, the bounding boxes usually do not completely surround the entire shape of the objects. One way to overcome this problem is to expand the rectangle by 10 to 20 percent and then show the window to the eight HOG detectors to identify whether it is a truck and determine its trajectory. It should be mentioned that the test window for the HOG classifier is cropped from the original image, not from the blob image. Figure 12 illustrates the process of this algorithm.

It is possible to send detected moving objects in every frame for classification, but the main issue is to maintain the real-time video stream. Because this recognition method is only applicable to videos and all of our test videos had 640x480 dimensions, we were able to
calculate runtimes for this size. The recognition computation time depends on number and sizes of segmented blobs, hence varied processing times between 0.01 to 1.7 seconds were achieved on 2.93 GHz processor. As a result, the system was set to recognize blobs obtained in every two seconds to process the videos without delay. This algorithm was tested on the same videos as before except that two of the 17 videos with a total length of 3:27 minutes were excluded because the viewpoint of the camera changed while recording. As noted, the Blob algorithm requires that the camera remain stationary to facilitate the foreground-background segmentation process. The ROC curves of the Blob-HOG approach along with Haar-HOG method are shown in Figure 10. This method could also process all the test videos with no delay.

**Discussion**

A similar research used HOG method to classify orientation of coarse vehicles (Rybski et al. 2010) which achieved 88% classification accuracy rate. Other study employed Haar and HOG fusion to extract features and could detect 94% of the cars (Negri et al. 2008). But there are significant differences between our research and those ones where the results of this research were achieved in a more complicated dataset. Firstly, the target of this research is to distinguish dump trucks from other equipment while the other researches just focused on vehicle detection regardless of their type. Secondly, earthwork operation fields are harsh and visually noisy environment compared to a typical urban environment and access roads, and loading and dumping zones are not as clear as urban streets where some dump trucks in our test sample were partially masked with soil depots, dust or emitted smoke from machine’s engine. Moreover, our test samples had wide range of camera angles where some images were taken from below, some from ground and some from higher levels which increased the level of complexity; however, comparing researches only tested the images from ground view. Finally, the second research (Negri et al. 2008) only focused on detection of the cars in front of driver (from rear view), which is more a straightforward problem than our case.

The Haar-HOG method slightly outperformed the Blob-HOG detection method. Imperfect performance of the foreground-background segmentation process was the main cause of the weaker performance. Moreover, the Blob-HOG approach suffers from three main problems. First, foreground-background segmentation can only be applied to videos where the camera remains stationary. If the viewpoint of the camera changes, the filter cannot process the video stream and will crash. Second, the foreground-background segmentation may absorb foreground objects if they stay motionless for a long time (Li et al. 2003). The last weak point arises from a deficiency in the segmentation filter, which does not always separate the entire moving object from background and sometimes segments a part of the moving object or may split one moving object into two or more disconnected pieces. Importantly, the Haar-HOG method can be used with moving and time-lapse digital videos, which are common ways of monitoring and documenting construction activity.

Despite these inefficiencies, Blob-HOG is less computationally intensive because it only triggers the costly object recognition engine if a moving object is detected. The Blob detection algorithm uses about 20% of the 2.93 GHz processor’s capacity for foreground-background segmentation, and when a moving object is detected, the usage increases to
more than 50%. In the Haar-HOG method, the system continually scans the video frames in the regular time intervals regardless of the existence of dump trucks in video scenes, which increases usage to more than 50% during scan intervals.

Both of the methods developed in this research have some advantages over existing recognition algorithms including detection based on color space, and background subtraction with Bayes or neural network classifiers (Chi and Caldas 2011a). First, color-based detection (Zou and Kim 2007) can segment the machines based on their color in plain backgrounds such as soil and snow, but because it uses color for recognition, it cannot perform robustly in a congested earthmoving site where different types of plants with similar color histograms are in the camera view. However, both Haar-HOG and Blob-HOG are invariant to color characteristics of the equipment and can robustly locate the machines among those with similar color as demonstrated in Figure 4 for instance.

Second, unlike the background subtraction plus normal Bayes or neural network classifiers (Chi and Caldas 2011a), which can only process movies from stationary cameras, Haar-HOG can also process images, time-lapse, and moving videos. Moreover, normal Bayes or neural network based detectors have been shown to classify only a limited number of trained objects. In that research (Chi and Caldas 2011a), the systems were tested on images containing combinations of only three semantic objects, namely skid steer loader, backhoe, and worker. Although it classified them with 96% accuracy (which is higher than 91% detection rate of this research) the algorithm has difficulty recognizing the target among unknown moving objects that are common in construction sites. The test videos used in this research, however, contained twelve types of moving objects including loaders, bulldozers, rollers, graders, pickups, SUVs, hydraulic excavators, water tankers, workers, truck mixers, tractors, and mobile concrete pumps and the HOG-based recognition algorithms could locate the dump truck regardless of other moving objects with 91% detection rate and 0.24 false alarms per frame.

Normal Bayes or neural network based classifiers used four features including aspect ratio (height/width), height-normalized area size, percentage of occupancy of the bounding box, and average gray-scaled color of the area, which work well in the detection of objects with apparently different shapes such as worker, mini loader and backhoe, but it is difficult for this algorithm to distinguish similarly shaped equipment such as a water tanker and a dump truck. The HOG based methods successfully distinguished the dump trucks from similar machines.

Applications
The recognition algorithms developed herein are now being used as a basis for an automated productivity measurement and performance control system. The next step to develop these envisioned systems is to pass detected objects in a form of bounding boxes to a tracking engine to extract the required data. A valid question is that why both of the presented frameworks continually scan images while once an object has been detected it only needs to be tracked (instead of detected again). There are two reasons for it which firstly it is probable to miss one or more target object at first scan, and secondly, dump trucks are highly dynamic machines and they frequently enter to camera view and exit it.
So it is necessary to repeatedly search the scene for the trucks and track with proper tracking algorithms. In addition, it is very straightforward to avoid double detection of an already detected and under track machines by checking whether the newly detected machines overlap with tracking ones. Because these systems are currently under development, detailed information cannot be presented in this paper; however, brief explanation is provided below to show the potential applications.

**Productivity measurement**

If the camera view captures the entryway to a loading or dumping area, it is possible to count the number of truck loads. For example, Figure 13 shows series of frames in 4.7 seconds intervals with a view of an access road from a rock quarry to a rock-fill dam construction site. The trucks toward the right haul rocks to the dumping zone, and the ones toward the left come back to loading area. The length of tested video was two minutes and eight seconds, and in total three loaded and two empty trucks passed the view. All of the passing trucks were identified by using Haar-HOG method, but one false positive was also observed in the processed frames.

![Figure 13. Counting the number of truck loads in about five seconds intervals (a - d)](image)

This short experiment was implemented by only using detection system. The next step is to pass every recognized machine to the tracking algorithms to avoid multiple counting of an already detected dump truck, and to track individual vehicles through their travel route.
**Pro-active real time control**

Automated equipment recognition is a practical tool to detect bottlenecks in an operation to allow for timely corrections. In large earthmoving sites such as earth-fill dams, highway construction, and surface mining operations, there are typically several loading units, such as loaders and excavators, working concurrently at different locations within the site. Trucks are assigned to these loading equipment based on optimization, simulation or other quantitative methods; however, due to numerous reasons such as human factors and equipment breakdowns, long queues may form in some loading zones. The models developed here may be used to detect, count and measure the waiting time of the trucks and inform the site superintendent if the number of waiting trucks exceeds a threshold. Figure 14 depicts excavation of the right abutment of a dam where five of six waiting dump trucks for a loader are correctly identified. The sixth truck in the far right of the image is missed.

![Figure 14](image)

**Figure 3. Queue of waiting dump trucks for a loader**

In addition to the mentioned applications, detection algorithms can be used for automated image and video indexing and retrieval. Captured images and videos may be scanned by a recognition system, allowing the detected objects to be tagged with the media file. It may also be used for safety management (Chi and Caldas 2011b).

**Conclusion and future directions**

Automated vision-based equipment recognition can be beneficial for several construction engineering applications, such productivity measurement and improvement, work-zone safety, locating resources in the site, and automated image and video tagging. This paper
presented two promising cascade approaches combining available image and video processing methods for detecting dump trucks in videos: Haar-HOG and Blob-HOG. In both methods, the HOG detector was the main classifier due to its robust performance; however, this algorithm is too slow for real-time applications. Thus two fast detection filters were added beforehand to limit the search areas for the HOG classifier. The Haar-HOG algorithm applies the Haar detection algorithm and Blob-HOG uses a foreground object detector to pass moving objects to the final classifier. The Haar-HOG method performed better with higher detection rates and lower false positives. It is also capable of searching for objects in videos regardless of the camera’s movement.

The next step is to merge these recognition frameworks with a tracking engine to extract the location of the machine at any point of time and to develop the mentioned applications. Although these algorithms were found suitable for rigid equipment, they did not provide promising results in the detection of articulated machines such as hydraulic excavators and loaders. Future work will focus on the development of a method to recognize equipment with deformable parts.

References


