This paper is published as:


Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos

Ehsan Rezazadeh Azar
Department of Civil Engineering, University of Toronto, Toronto, ON, Canada, M5S 1A4
Corresponding author, Tel: +1-416-8317631, Fax: +1-416-978-5054
rezazade@ecf.utoronto.ca

Brenda McCabe
Department of Civil Engineering, University of Toronto, Toronto, ON, Canada, M5S 1A4
brenda.mccabe@utoronto.ca

Abstract

Detection of earthmoving equipment in construction images and videos can increase the automation level of many construction management tasks such as productivity measurement, locating of machines, work-zone safety, and semantic image and video indexing. Some of the earthmoving plants, such as hydraulic excavator; have articulated shapes making them a difficult target for even state of the art object recognition algorithms. The goal of this paper is to develop a model for non-rigid equipment detection, and pose estimation in construction images and videos. In this paper, we describe an object recognition system based on mixture of appearances of deformable body parts of the hydraulic excavator and compare its results with general histogram of oriented gradients detectors in both images and videos. Then a spatial-temporal reasoning model is presented which uses time and space constraints of the excavators’ moving patterns to improve the detection results in videos.

Key Words:
Computer vision; Object recognition; Earthmoving equipment; Automated data collection; Digital images; Spatial-temporal reasoning.

1 Introduction

A variety of earthmoving equipment is used in heavy civil construction. Many of these plants are manufactured specifically to carry out one type of operation, while others, such as hydraulic excavators, can handle multiple activities, in this case excavation, loading, trimming and moving materials. To improve the productivity of these costly resources, contractors generally try to minimize their idle time and non-value adding activities [1]. While smaller projects can be effectively managed using experienced site managers, larger projects benefit from the use of simulation or other methods to optimize resource utilization and therefore productivity. However, planning methods such as simulation, linear programming, and genetic algorithms do not always accurately predict the actual production rates. Human factors, equipment breakdowns,
continually changing travel routes, degradation of the quality of site roads, and weather conditions are a few of the factors affecting the earthmoving operations. As a result, there has been growing interest in methods to monitor the operations in real-time to find discrepancies between the planned and real performance, record productivity data for future projects, and to make timely changes to improve the operations as quickly as possible [2]. Automated vehicle tracking technologies such as global positioning system (GPS) [3, 4] and ultra wideband (UWB) [5] have been used to automate data collection. These real-time positioning devices provide the three-dimensional location of the equipment, which is useful to interpret the activities of mobile machines such as dump trucks and rollers [3, 4]; even though they cannot effectively distinguish productive activities from non value-added (NVA) traverses. These tools can also locate the stationary plants such as excavators, which is valuable data for planning and management of these resources; however, they cannot distinguish whether the equipment is idle or active.

Various built-in sensing devices can provide a wide range of data from the machine itself such as engine operating parameters [6], body, boom, and bucket orientations [7] to facilitate operating the equipment efficiently. It is also possible to collect and interpret these data to measure the productivity of the machine. However, there are some limitations with these devices related to cost-effectiveness, and data interpretation. In addition, all of the mentioned devices have issues in rented equipment, which is commonly employed by general contractors, because of the effort and cost to continually install and remove sensing tags from the equipment and update the monitoring software. Leaving technology aside, manual optimization through continuous supervision of the operation is labour-intensive, expensive, and error prone.

Computer-assisted visual monitoring has potential to detect, track and measure the productivity of stationary and mobile equipment. Because major projects generally take place in open field and mining sites, sight lines can be selected that are not obscured. Although many construction sites already have surveillance cameras that capture videos or time-lapsed photos at regular time intervals [1, 8], they are not often used to their full potential due to the labour-intensive process of manually extracting data from the images and videos before they are deleted in a rotation to save memory. Computer vision algorithms are now being evaluated to automate data extraction from construction videos for tracking resources and personnel [9, 10], measuring concrete pouring cycles [8], and assessing human workforce productivity [11, 12].

This paper describes part of an extensive research effort to develop an automated vision-based system to monitor earthmoving activities. The envisioned system would have three main functions: object detection, object tracking, and activity recognition. The goal of object recognition is to find and identify the equipment on site from video frames. Existing object detection algorithms such as Histogram of Oriented Gradients (HOG) [13] show promising performance in recognition of rigid objects such as urban vehicles [14] and off-highway dump trucks as an example of rigid construction equipment [15]; however, deformable object detection and articulated pose estimation are more challenging problems faced by the computer vision community [16, 17]. A number of earthmoving equipment have non-rigid frames that move or deform their parts to perform the work, including hydraulic excavators, loaders, and articulated dump trucks.
Hydraulic excavators are highly articulated and are extensively used in construction due to their versatility to carry out a variety of operations in different environments from quarries to compact urban sites. They have also been the subject of other research such as automating excavator operation [18], and improving equipment safety [19]. Measuring the idle time of hydraulic excavators using color space values [1] to isolate the machine in images with soil and snow backgrounds was attempted. In another research, moving entities in construction videos were segmented by a background subtraction filter, and the remaining blobs were classified with respect to four features including aspect ratio, height-normalized area size (occupied area size in pixels/ centroid of the area height), percentage of occupancy of the bounding box, and average gray-scaled color of the area [20].

In this paper we firstly describe the cutting edge methods currently used to detect deformable objects, and then present our part-based recognition algorithm, which was inspired by those methods. Afterwards, we employ spatial-temporal constraints of the moving patterns of the excavators to eliminate false detections and validate true positives. Finally, the evaluation results of the algorithms on a test dataset are presented.

2 Articulated object recognition

Object recognition in digital images is challenging due to changing viewpoints, illumination, occlusions, and scale. Robust algorithms have been developed in the last decade to detect rigid objects such as Histograms of Oriented Gradients (HOG) [13] and Haar-like features [21]. In HOG algorithm, which was one of the most successful rigid object detection methods, computed gradients of the gray-scaled image are discretized into spatial and orientation cells to form a descriptor vector. Then positive and negative vectors from a training dataset are trained by the SVMLight method [22] resulting in a single classifying vector. This detector uses a sliding window technique to check all locations and scales of an image. If the result exceeds a determined threshold, that window is accepted as a target.

Dramatically more difficult to detect are objects that change shape, and current research is focused on human detection and pose identification due to the complicated configurations of the human body. The outcomes are highly applicable for security surveillance, traffic safety, and image indexing. Many of the recently developed models use part-based and pictorial structures to find a group of parts of a semantic object arranged in a deformable configuration [16, 17]. One of the cutting edge algorithms is latent support vector machine (Latent SVM) part based models [17], which won the 2009 PASCAL object detection challenge [23]. This model uses a modified HOG detector, called a root filter, to locate the most probable candidates for the object within the image. It then searches for the parts of the object at twice the spatial resolution relative to the features captured by the root filter inside the detected root areas. In the current research, we used a similar idea with substantial modifications to detect a root and then search for the possible configurations of the parts of the excavator to both detect and estimate the pose of the equipment.

3 Research methods

3.1 Deformable parts

Hydraulic excavators are highly deformable machines that can slew 360 degrees and rotate all three parts of their arm (boom, dipper, and attachment) around their hinged supports. The typical
deformations of the machine are illustrated in Fig. 1. As a result, the machine can have countless forms, making it impossible to be detected with a limited number training configurations as used in the case of rigid-frame equipment [14, 15].

![Figure 1. Deformations of the hydraulic excavators](image)

In latent SVM part-based models [17], the first classifier is trained to detect the entire body (e.g. human), which is called the root. Then it searches for the body parts (e.g. torso, arms, and legs) inside the root to validate the detection. An excavator can have multiple shapes with parts of the machine masked with soil deposits or by other equipment in the camera view. In early testing, it was very difficult to find the root candidates (an excavator) with one or two root classifiers, so the approach was modified. Instead of defining the root as the whole body of the object, a major part of the body that is most visible was used as a root. Then instead of searching within the root for identifiable parts, the algorithm was set up to search for the adjacent parts in a variety of possible configurations to finalize the recognition process. The first part of the arm, called the boom, was selected as the root (See Fig. 2) and the dipper (second section of the articulated arm) as the adjacent part. Neither the main body nor the bucket was considered as the root or adjacent part for several reasons. First, except for long boom excavators, the boom and dipper have approximately similar shapes and size ratios in across different makes and sizes of excavators. Body and cabin forms in the market; however, can differ substantially. For example, urban excavators have very compact bodies that enable them to operate in confined working areas and minimize swing collisions. Second, not all of the excavators have the same attachments at the end of their dipper; for example, they can carry pneumatic hammers, buckets, trenchers or other types of add-ons to perform specific operations. Identifying attachments will be of interest in the future as a means to help identify the activity of the equipment. For the purposes of this paper, however, we will focus on the more consistent features of the excavator.

The detector for the root (boom) was trained in left and right configurations. Some of the training samples are shown in Fig. 3. Because the dipper can rotate around the hinge with the boom, the detector was trained for the six views shown in Fig. 4. Based on these body parts, an excavator can have six poses including left-horizontal, left-inclined, left-vertical, right-horizontal, right-inclined, and right-vertical. In complete front and rear configurations where the boom is aligned with camera view, it is impossible to distinguish the parts and we need to train separate detectors; those poses were not considered for this research.
3.2 Features

The HOG method was used to train a classifier for each individual component of the object. Fig. 5 shows a sample excavator arm and the visualization of its HOG features. In total, eight classifiers including two roots and six parts are trained and Table 1 shows the statistics of the positive and negative samples used to train each detector. These images were collected from a multimedia archive of a construction company, publicly available online sources, and photos manually captured by one of the authors from two earthmoving projects including a rock-fill dam construction project and excavation of the foundation of a high-rise building.
Figure 4. Views of the dipper

Figure 5. Left: a sample excavator arm, right: HOG features are extracted

<table>
<thead>
<tr>
<th>Part</th>
<th>Root-left</th>
<th>Root-right</th>
<th>Horizontal-right</th>
<th>Inclined-right</th>
<th>Vertical-right</th>
<th>Horizontal-left</th>
<th>Inclined-left</th>
<th>Vertical-left</th>
</tr>
</thead>
<tbody>
<tr>
<td># Positive samples</td>
<td>1040</td>
<td>800</td>
<td>398</td>
<td>791</td>
<td>926</td>
<td>398</td>
<td>791</td>
<td>926</td>
</tr>
<tr>
<td># Negative samples</td>
<td>7700</td>
<td>7700</td>
<td>7700</td>
<td>7700</td>
<td>7700</td>
<td>7700</td>
<td>7700</td>
<td>7700</td>
</tr>
</tbody>
</table>

Altogether 770 negative images were collected as the base negative example training set. These images contained construction landscapes and earhmoving equipment other than hydraulic excavators, such as trucks and loaders, to decrease the chance of misclassification of those machines. Ten windows were randomly cropped from each image and scaled to corresponding viewpoint sizes, which resulted in 7700 negative training samples for each part. Some negative training samples are illustrated in Fig. 6. HOG training is a two-round process in which the
initially trained classifier searches the original negative images (770 images in our case) and all of the detected windows, called hard negatives, were scaled and added to the cropped negative images for the second round of training. Open source OpenCV 2.1 [24] library was used to calculate the HOG descriptors and publicly available SVMlight [22] software was utilized to train the classifiers from the calculated descriptors.

Figure 6. Negative training samples

3.3 Mixture models

A mixture model with n pieces can be described by n-tuple, $P = (P_1, \ldots, P_m)$, where $P_i$ is the i-th part of the articulated object, which in our case included two parts ($m=2$), the root and the adjacent part. Each of the components has a possible location and a HOG descriptor. As depicted in Fig. 7, this recognition model is a two stage detection process with both implemented using a multi scale HOG detector [13].

Figure 7. Flowchart of the part-based detection algorithm
The first step of the detection process is to search the image for two directions (left, right) of the root, which may produce several candidate boxes. Then the system searches for the dipper in possible regions adjacent to the detected roots. For example, if the root detector finds a root in “left” orientation, then the dipper must be in the left side of the boom in one of the three possible configurations including “left-horizontal”, “left-inclined”, and “left-vertical” as illustrated in Fig. 8. The sizes of these search areas are based on the size of the detected root as presented in Table 2. Various size ratios were examined to obtain the best detection rate and run-time efficiency. If the region size increases, it will amplify the computation time and increase the possibility of false detection. On the other hand, smaller search areas may miss the dipper. The selected dimension ratios are large enough to contain all of the regular dippers, except for long boom excavators, and maintain the run-time efficiency of the algorithm. This recognition software was coded using C++ language, and OpenCv 2.1 [24] libraries were used for image and movie processing purposes.

<table>
<thead>
<tr>
<th>Search regions</th>
<th>Width</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horizontal</td>
<td>= width of the root</td>
<td>= 0.7 height of the root</td>
</tr>
<tr>
<td>Inclined</td>
<td>= width of the root</td>
<td>= 1.2 height of the root</td>
</tr>
<tr>
<td>Vertical</td>
<td>= width of the root</td>
<td>= 1.4 height of the root</td>
</tr>
</tbody>
</table>

Figure 8. Possible search regions for each configuration of the dipper

4 Empirical results

4.1 Static images

A valid question here is that whether it is necessary to use part-based model or just utilizing general HOG detectors can resolve the problem. Therefore, two sets of experiments were carried out to figure out if the part-based method can improve the results in compare to standalone HOG detectors. In the first set of experiments, only the root detectors including left and right orientations were used and for the second test, the whole part-based algorithm was examined. These models were evaluated using 253 images of different sizes, containing 284 excavators varying in make and pose. These images were randomly selected from the collected image dataset; none of them were used for training of the classifiers. These photos were taken in busy construction sites such that many of the images contained other types of equipment in addition to
the hydraulic excavators, allowing the capability of the algorithm to correctly recognize the hydraulic excavator to also be evaluated.

These images were shown to detectors with different thresholds, which alter the results. As the threshold increases, both of the detection rate and false alarms decrease. A true positive requires that a detected bounding box not only overlap more than 50% with a ground-truth bounding box as stated in the evaluation rules of PASCAL visual object classes challenge [25], but also should correctly estimate the pose of the excavator. In addition, multiple detections of the same machine in one image are penalized. Table 3 presents the results of two methods together where the thresholds of the root detectors are identical in the same rows. To have a better overview of the results, the receiver operating characteristic (ROC) curves of the two different approaches are plotted together and shown in Fig. 9.

Table 3. Results of the General HOG and part-based methods

<table>
<thead>
<tr>
<th></th>
<th>General HOG (only roots)</th>
<th>Part-based method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection rate (%)</td>
<td>False positives per frame</td>
<td>Detection rate (%)</td>
</tr>
<tr>
<td>Test 1</td>
<td>61.62%</td>
<td>0.27</td>
</tr>
<tr>
<td>Test 2</td>
<td>72.18%</td>
<td>0.52</td>
</tr>
<tr>
<td>Test 3</td>
<td>77.46%</td>
<td>0.99</td>
</tr>
<tr>
<td>Test 4</td>
<td>81.34%</td>
<td>1.81</td>
</tr>
<tr>
<td>Test 5</td>
<td>85.56%</td>
<td>3.60</td>
</tr>
</tbody>
</table>

A comparison of the two methods with the same thresholds in Table 3 shows that part-based method can significantly reduce false detections; however, it decreases the detection rate as well. So the pair wise comparison does not provide a reliable conclusion, and the overall performance of the two approaches should be considered to compare the result at any given detection rate or false alarm. ROC curves are the best means to find differences. As illustrated in Fig. 9, both methods have almost the same detection rate in the range of 0.52 to 2.19 false alarms per frame, but the main difference is in the ranges of lower than 0.52 false alarms per frame where the part-based method considerably outperforms the general HOG method. In fact, this range is the most important choice for video recognition applications such as productivity measurement or safety management, where the cost of misclassification is high and it is necessary to avoid false detections as much as possible while maintaining an acceptable detection rate. In addition, the
part-based method also estimates the orientation of the dipper, which is useful data for activity recognition.

Fig. 10 depicts some detections achieved using the part-based method. Because the general HOG method is invariant to illumination and scale, the part-based method (which uses HOG descriptors) also showed good performance in detection of various sizes of excavators with different colors and illuminations. For example, Fig. 10c shows a photo taken in sunset time with very low lighting where the excavator is detected successfully, while the other samples in Fig. 10 are taken in mild (Fig. 10d) to very bright conditions (Fig. 10a).

![Figure 10. Detection samples](image)

Although both methods demonstrated considerable performance in detection of the excavator in the scenes where the arm of the machine is visible, they fail to recognize the equipment if the arm is not noticeable or is aligned with the camera view. Most of the false positives occurred in the correct locations but with wrong boom direction. For instance, the detector spotted the excavator in the “right-inclined” configuration in addition to “left-inclined” (see the left machine in Fig. 10b). Another apparent problem in the part-based method was observed in detection of the dipper, where the system recognized the dipper in two different poses at the same time. There were some instances that the secondary classifier detected the dipper in both inclined and vertical, or horizontal and inclined configurations. The main reason is due to overlapping of the samples in the training stage. The samples were manually divided into categories and there were some misclassified samples due to human error. The possible remedy for this issue is to calculate the ratio of the detection score over the threshold; then in the case of the overlapping detection, select the one that has the higher number as we did for this research.

The run-time efficiency of both algorithms showed fairly good results; it takes less than seven seconds to process a low resolution standard 640x480 pixel frame on a 2.93 GHz CPU. Table 4 shows process runtimes on different image sizes on the same CPU. General HOG detectors have
stable process times on images of the same size, while the part-based method showed varied runtimes in those images. The part-based process times changed because of the varied number of secondary search regions generated by the root detectors.

<table>
<thead>
<tr>
<th>Image size (Pixels)</th>
<th>General HOG (Sec)</th>
<th>Part-based (Sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>640x480</td>
<td>6.5</td>
<td>6.5-7.0</td>
</tr>
<tr>
<td>1024x768</td>
<td>17.1</td>
<td>17.1-20</td>
</tr>
<tr>
<td>1920x1080</td>
<td>45.9</td>
<td>45.9-51</td>
</tr>
<tr>
<td>2592x1944</td>
<td>116</td>
<td>116-120</td>
</tr>
</tbody>
</table>

Being able to both locate and estimate the pose of the articulated equipment can add functionality to the automated monitoring systems. The pose of the excavator is an indicator of the working state of the machine. For instance, Fig. 11 shows two dump trucks close to a working excavator where one of them is being loaded and the other is waiting. This image was shown to the part-based classifier and the rigid equipment recognition algorithm [15] to detect excavators and dump trucks respectively. In the 2D view, both of the trucks are close to the working excavator, making it difficult for the system to distinguish between the waiting and loading trucks. However, the pose of the excavator is additional information that can be used to indicate that the truck being loaded is the one nearest the dipper in the “inclined” pose.

Figure 11. Application of the excavator poses to interpret the interactions

4.2 Videos

The aim of this research is to detect hydraulic excavators in construction videos so that the detected objects can be passed to the tracking engine for further analysis. However, the scope of this paper is limited to the recognition of excavators in the videos. Hydraulic excavators are generally stationary and only move to change their working zones, so a recognition system needs to detect them only once in videos. We recorded 21 test videos with a total duration of two hours.
and twelve minutes from three construction projects. All of these videos had 640x480 pixel resolution.

The system scans the videos in ten seconds intervals until it detects an object regardless of being true of false. As shown in Table 4, the maximum processing time for 640x480 frames is seven seconds, so ten seconds time interval is adequate to maintain real-time monitoring of the video. Both the general HOG and part-based methods were evaluated with different thresholds to compare their performance. There are two evaluation criteria for this test: detection rate and the average time to find the first object in the videos. The results are presented in Table 5.

The tests with lower ID numbers have higher thresholds resulting in higher precision, and as the ID number increases, the threshold decreases. The test cases with higher thresholds have higher precision, but it takes longer (more search frames) for them to detect the first object. As the threshold decreases, the system finds objects faster at the cost of more false alarms. As hypothesized in the previous section about the part-based algorithm outperforming at higher threshold tests, the part-based outperformed the standalone HOG method in the test videos, even though it took longer and more search windows to find the first object.

5 Spatial-temporal reasoning

Yet, the existing object recognition algorithms are far from ideal, and they have a long way to go to reach the ultimate target of matching human vision [26]. Even the most advanced methods have type one and type two errors. Excavators are stationary, so it is not necessary to search for them in predetermined time intervals. Once they are detected, their positions are passed to the tracking engine. Thus, there is only a single chance for detection and the cost of false positives can be very high. As shown in Table 5, detectors with higher thresholds need several search frames to find the machine, and in addition, there is no guarantee for 100% correct detection. Spatial-temporal reasoning is a new concept in computer science that employs commonsense background knowledge to interpret situations. In contrast to mathematical and artificial intelligence theories, it uses inexpensive constraint analysis including reasoning about entities located in space and time [27]. As a result, logical reasoning has been employed for object and movement recognition in image sequences [28, 29].

<table>
<thead>
<tr>
<th>Method</th>
<th>Test ID</th>
<th>Detection Rate</th>
<th>No. of False Detections</th>
<th>First detection (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General HOG</strong></td>
<td>HOG0</td>
<td>85.71%</td>
<td>2</td>
<td>53.33</td>
</tr>
<tr>
<td></td>
<td>HOG1</td>
<td>76.19%</td>
<td>5</td>
<td>38.10</td>
</tr>
<tr>
<td></td>
<td>HOG2</td>
<td>71.43%</td>
<td>6</td>
<td>29.05</td>
</tr>
<tr>
<td></td>
<td>HOG3</td>
<td>71.43%</td>
<td>12</td>
<td>12.86</td>
</tr>
<tr>
<td></td>
<td>HOG4</td>
<td>61.90%</td>
<td>24</td>
<td>11.90</td>
</tr>
<tr>
<td><strong>Part-based method</strong></td>
<td>P-B0</td>
<td>90.48%</td>
<td>0</td>
<td>90.00</td>
</tr>
<tr>
<td></td>
<td>P-B1</td>
<td>90.48%</td>
<td>1</td>
<td>64.29</td>
</tr>
<tr>
<td></td>
<td>P-B2</td>
<td>90.48%</td>
<td>2</td>
<td>35.71</td>
</tr>
<tr>
<td></td>
<td>P-B3</td>
<td>76.19%</td>
<td>8</td>
<td>14.76</td>
</tr>
<tr>
<td></td>
<td>P-B4</td>
<td>76.19%</td>
<td>10</td>
<td>13.33</td>
</tr>
</tbody>
</table>
A spatial-temporal reasoning algorithm was developed for this research to improve the detection rate in videos. In this framework, the HOG thresholds are set low to produce multiple detections including true positives and false alarms. Although several false positives are detected, the risk of false negatives is virtually eliminated. The system searches 10 frames in the first minute of the movie (one every six seconds) for a hydraulic excavator, and then the detected boxes in the frames are grouped based on defined spatial-temporal constraints. These constraints include consistency of sizes, displacement, and directions (left or right) of the detected boxes. Two movies of working excavators with a total length of twenty-one minutes were studied to determine these constraints. One frame in every six seconds was scanned and the true positives were carefully examined to extract the constraints, and the outcomes are presented in Table 6.

Detection boxes around objects are called “nodes”, and groups of similar nodes are called “paths”. This algorithm first finds an object, and then searches the next video frame for a similar object that satisfies the constraints. Even if a node is drawn in the next frame, the system loops through all of the remaining frames for matches. Two nodes of a path cannot be in the same frame. Fig. 12 shows nodes created in the first minute (images a through j) and frames k to n in Fig. 12 illustrate the four detected paths.

Figure 12. Object recognition in time intervals (images a to j), and four detected paths (images k to n)
Afterwards, the paths are evaluated based on logical movements of an excavator to distinguish the path of an excavator from others. This process involves another spatial-temporal reasoning by selecting the path closest to the movement pattern of an excavator. Again the same two videos were investigated to develop the reasoning.

As observed in the test videos, there are two types of false positive paths. The first type includes paths with one node, which results from scattered false detections (frame n in Fig. 12). The second type of false alarm path is due to repetitive misclassification within the same region in several frames (frame l and m in Fig. 12).

<table>
<thead>
<tr>
<th>Table 6. Spatial-temporal constraints of the true positives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constraint</strong></td>
</tr>
<tr>
<td>Size</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Displacement</td>
</tr>
<tr>
<td>Direction</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

On the other hand, the paths of true positives contain a cluster of boxes with some displacement, small size variations, and logical changes in direction of the boom due to movements of the working excavator (frame k in Fig. 12).

To select the correct path, the system first sorts the paths based on the number of nodes, because the true path representing a working excavator is always amongst those with the highest number of nodes. The candidate paths with few nodes usually contain the repetitive false positives. So the path with higher change of directions and displacements is selected as the target object. This framework processed the first sixty seconds of the 21 test videos and correctly recognized twenty out of twenty one machines (95.2% detection rate) with just one false detection. This method was evaluated on scenes with one working excavator, and further modifications are required to make the system applicable for detection of multi excavators in video streams. This method is only applicable in the videos with static backgrounds, because the reasoning about the location of the detected boxes is possible only in the movies captured by stationary cameras.

The spatial-temporal framework of this research has some advantages over existing recognition algorithms including color-based detection [1], and foreground-background segmentation with Bayes or neural network classifiers [20], which have an accuracy rate (96%) very close to the outcome of this research (95.2%). First, color-based recognition [1] can find the machines based on their color in plain backgrounds such as soil and snow; however, the system has difficulty in finding the excavator in congested sites with different types of equipment, particularly where they were painted with similar colors. Second, unlike the normal Bayes or neural network based detectors which can classify only a limited number of trained objects, HOG based detectors can locate the target regardless of other existing objects. In that research [20], test videos contained only three types of objects including a mini loader, backhoe, and workers and the system had the
corresponding classifiers; however, test videos in this research contained several moving objects including bulldozers, rollers, pickups, SUVs, dump trucks, workers, truck mixers, pile driving machines, and mobile concrete pumps.

6 Conclusion

A part-based object recognition model is introduced that uses discriminately trained HOG classifiers to detect hydraulic excavators in different poses. This method first scans the image for the first part of the boom and then searches for the second part in possible configurations around the boom to validate the detection and estimate the pose of the machine. This algorithm showed almost as the same performance as general HOG classifiers, but it noticeably outperformed the HOG detectors in higher thresholds where it had higher detection rates with lower false alarms. Both of these methods had costly misclassifications in videos, thus the moving patterns of the machine were translated to involve spatial-temporal constraints to eliminate false alarms and keep the cluster of true positives. This combination showed very promising results in recognition of a single excavator in videos from stationary cameras. Future work will focus on the integration of object detection and tracking algorithms to recognize the activities and measure the productivity of the hydraulic excavators.

7 References


