Automation in Construction 32 (2013) 24–37

Contents lists available at SciVerse ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors

Milad Memarzadeh [a,1], Mani Golparvar-Fard [b,*], Juan Carlos Niebles [c,2]

[a] Vecellio Construction Engineering and Management, Via Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, USA
[b] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[c] Electrical and Electronic Engineering, Universidad del Norte, Colombia

ARTICLE INFO

Article history:
Accepted 4 December 2012
Available online 1 March 2013

Keywords:
Histogram of oriented gradients
Support vector machine
HSV colors
Resource tracking
Performance monitoring

ABSTRACT

This paper presents a computer vision based algorithm for automated 2D detection of construction workers and equipment from site video streams. The state-of-the-art research proposes semi-automated detection methods for tracking of construction workers and equipment. Considering the number of active equipment and workers on jobsites and their frequency of appearance in a camera's field of view, application of semi-automated techniques can be time-consuming. To address this limitation, a new algorithm based on Histograms of Oriented Gradients and Colors (HOG + C) is proposed. Our proposed detector uses a single sliding window at multiple scales to identify the potential candidates for the location of equipment and workers in 2D. Each detection window is first divided into small spatial regions and then the gradient orientations and hue–saturation colors are locally histogrammed and concatenated to form the HOG + C descriptors. Tiling the sliding detection window with a dense and overlapping grid of formed descriptors and using a binary Support Vector Machine (SVM) classifier for each resource enables automated 2D detection of workers and equipment. A new comprehensive benchmark dataset containing over 8000 annotated video frames including equipment and workers from different construction projects is introduced. This dataset contains a large range of pose, scale, background, illumination, and occlusion variation. Our preliminary results on detection of standing workers, excavators and dump trucks with an average accuracy of 98.83%, 82.10%, and 84.88% respectively indicate the applicability of the proposed method for automated activity analysis of workers and equipment from single video cameras. Unlike other state-of-the-art algorithms in automated resource tracking, this method particularly detects idle resources and does not need manual or semi-automated initialization of the resource locations in 2D video frames. The experimental results and the perceived benefits of the proposed method are discussed in detail.

1. Introduction

Over the past few years, many construction companies have started online video streaming from their job sites. Detailed and continuous videos of the work-in-progress provide an excellent opportunity for automated performance assessment and enable timely identification of productivity, safety, and occupational health issues. Moreover, site video streams provide an ideal test bed for developing computer vision algorithms for automated performance assessment in dynamic construction conditions.

Despite all the benefits, to date application of these video streams for the purpose of automated activity analysis is still unexploited by researchers. A major reason is that these video streams are not in a form that is amenable for automated processing, at least by traditional computer vision methods. They are widely variable in terms of their location and field of view, have uncontrolled illuminations, resolution, and Pearson Education Inc. qualities. Most importantly, they consistently suffer from static and dynamic visual occlusions caused by the physical construction progress or movement of workers and equipment. Developing computer vision algorithms that can operate effectively with such video streams require 1) automated and real-time 2D detection of the equipment and workers from single cameras; 2) synchronized detections across multiple cameras and localization of the resources in 3D; and finally 3) automated action recognition. Within this scope, this paper focuses on the first key challenge, which is *automated 2D detection*; i.e., knowing what resources are visible within a camera's field of view and continuously track them for the entire period of time the resource is visible. Robust 2D detection provides an opportunity for continuous 3D localization and action recognition, which are critical components for any automated vision-based performance assessment system. While a number of researchers have looked into developing vision-based assessment methods (Section 2), many challenging problems remain open.

As a step towards fully automated performance assessment methods, this paper focuses on the problem of automated 2D detection of workers

[*] Corresponding author. Tel.: +1 217 300 5226; fax: +1 217 333 9464.
E-mail addresses: miladm@vt.edu (M. Memarzadeh), mgolpar@illinois.edu (M. Golparvar-Fard), njuan@uninorte.edu.co (J.C. Niebles).
[1] Tel.: +1 540 557 7087; fax: +1 540 231 7532.
[2] Tel.: +57 5 350 9270; fax: +1 540 231 7532.

0926-5805/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.autcon.2012.12.002

and equipment in site video streams and a number of applications this enables. Fig. 1 shows examples of the technical challenges associated with using video streams for 2D detection of excavators, dump trucks, and workers. Not having a priori knowledge about the appearance, pose, location, and scale of the resources makes the task of detection extremely difficult. Given fixed cameras with small lateral movements, cluttered background, moving equipment and workers with deformable body configurations, the task is to automatically and reliably detect and localize these dynamic resources in 2D.

As such, first the current practice of the industry and the state-of-the-art in research are overviewed. Next, a set of open research problems for the field are discussed, including automated detection of those resources that had previously left the camera's field of view and real-time tracking. The new method expands on the work originally presented in [9] with addition of several novel components to the algorithm that significantly improve the performance of the method. It is also accompanied with extensive validation experiments. A comprehensive dataset and a set of validation methods that can be used in the field for development and benchmarking of future algorithms are also provided. The perceived benefits and limitations of the proposed method in the form of open research challenges are presented. Videos of the proposed method, along with additional supplementary material can be found at http://www.raamac.cee.vt.edu/detectiontracking.

2. Background and related work

A large number of construction companies are still using traditional data collection methods for performance analysis including direct manual observations, methods adopted from stop-motion analysis in industrial engineering [34], and survey based methods. Although these methods provide beneficial solutions in terms of improving performance, their applications due to the large size of the data that needs to be collected are labor-intensive [22,39] and can be subjective [18]. The significant amount of information which needs to be collected may also adversely affect the quality of the analysis [17,20]. Such limitations minimize the opportunities for continuous benchmarking and monitoring which is a key element in productivity improvement [33]. In recent years, several researchers have focused on developing techniques that can automate the entire process of performance monitoring. These techniques mainly focus on tracking of construction workers and equipment as a critical step towards automation of performance assessment. In the following, these methods are reviewed and their limitations are discussed.

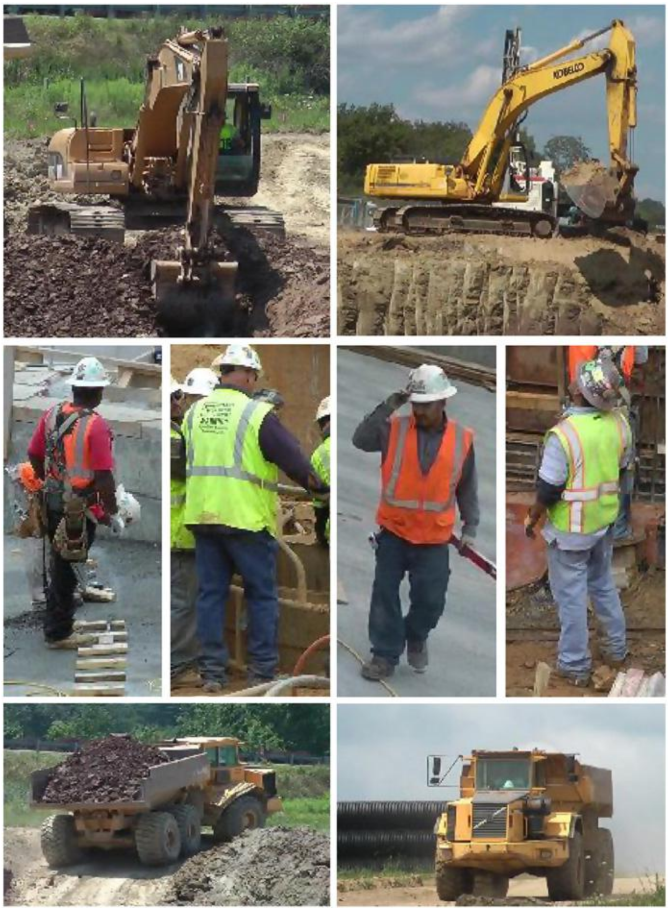

Fig. 1. Example frames from video sequences of excavator, truck and worker operations. Each row illustrates different body postures and configurations, which challenges development of automated 2D tracking methods.

2.1. 3D localization and tracking of construction resources using sensors

In recent years, a number of research studies [19,22,23,39] have focused on developing techniques to automatically localize and track construction resources in 3D. The main goal of these methods is to support improvement of operational efficiency/safety and, in turn, minimize idle times. To address this need, different tracking technologies such as barcodes and RFID tags [11,13,24,32,37,38], Ultra WideBand (UWB) [5,25], 3D range imaging cameras [21], global and local positioning systems (GPS) [21,25], and computer vision techniques [3,35] have been explored. Among these, UWB methods can detect time-of-flight of the radio frequency at various frequencies, which allows for providing 2D and 3D localizations even in the presence of severe multipath [15]. In a recent case, Teizer et al. [25] applied the UWB technology for real-time tracking of resource locations in 3D. This UWB system requires resources including the workers to be individually tagged and satisfactory positioning data to be transferred to the system prior to its implementation [3]. As such, the implementation of this system may be challenged and can be costly where they are hundreds of construction resources that need to be tracked. Recent research has focused on the use of 3D range imaging camera for spatial modeling [21] and resource tracking [25] on construction sites. The low resolution and short range of these cameras can challenge the application of these systems on large-scale construction sites.

GPS modules have also been used for positioning of equipment and surveying purposes [4]. Despite the wide range of benefits that GPS can offer to the construction industry, using it for tracking workers in interior spaces can be challenging. GPS mainly operates outdoors, and needs to be regularly attached to the resource that is being tracked. Consequently, tracking construction resources in particular workers with GPS can be infeasible in several cases. In the most recent research effort, an inertial measurement unit Personal Dead Reckoning (PDR) system which does not require pre-installed infrastructure is proposed [26]. This method is accurate for tracking workers outdoors. Nonetheless, its accuracy degrades with both path complexity and the time spent indoors. Once the accumulated drift exceeds the acceptable threshold, the user needs to step outdoors and recover the GPS signal to reset the system. More research needs to be done on application of such systems for continuous tracking purposes.

RFID tags have high durability in harsh environments, do not require line-of-sight, and can be embedded in concrete. Unless combined with other techniques, RFID can only function within a fixed radius inside which the resource exists [3]. As a result, several research studies [11,13,31,32] combined RFID with GPS technology for the purpose of automated localization and tracking of construction equipment. Despite the potential, RFID tags still require a comprehensive infrastructure to be installed on the jobsite, which can be very costly. The near-sightedness of RIFD also limits the applicability of real-time tracking, and due to GPS applications, the line-of-sight in many locations may adversely impact their benefits.

Although these techniques may accurately track location of the workers and equipment in 3D, yet do not provide information about the nature of the operation or the *actions* in which the workers or equipment are involved. Without information about these actions, performance cannot be measured automatically.

2.2. Vision based tracking and 3D localization of construction resources

Site video streams have long been used in the Architecture/Engineering/Construction (AEC) community for systematic activity analysis of site operations [34]. Compared to sensor-based approaches, videotaping is cost-effective and enables action recognition of construction resources. This is a key benefit for activity analysis and formation of crew-balance charts for craft productivity assessment purposes. Despite the popularity of onsite observations [12] or video-based activity analysis [34], these techniques are still primarily manual and involve tedious processes. As such, their applications for benchmarking and continuous assessments are not widely applied and are still limited to certain projects. Several recent studies [3], [16], [17], [22], [36] and [32], have emphasized on the need for automated video-based performance assessment techniques. Development of automated video-based methods for action recognition or 3D resource tracking first requires the workers and equipment to be detected in 2D. Recently developed methods [44] are either simulated in controlled environments or have primarily focused on automating the 3D tracking assuming semi-automated detection of resources in 2D which is the pre-requisite to 3D tracking. Others such as [3,35,42] use a priori knowledge for their assessments such as expected known locations for tracking tower crane [42], or application of Scale Invariant Feature Transforms (SIFT) [28] and Speeded Up Robust Features (SURF) [2] for initial recognition which limit the application of these methods for automated performance assessment.

Recent works [1], [36], and [6] focus on developing techniques for automated 2D detection and localization of construction workers and equipment. Particularly, Chi and Caldas [6] proposes a background subtraction algorithm to differentiate between the moving object and the stationary background and uses the Naïve Bayes and Artificial Neural Networks algorithms for learning and classification. Despite the good performance, the background subtraction component of their algorithm does not allow *idle* resources to be detected which can limit its application for productivity and resource proximity (safety) assessment purposes. Several existing object detection and background subtraction algorithms are combined and used for learning and 2D tracking of off-highway dump trucks in video streams [1]. Particularly, the application of HOG detectors [9], Haar-like detectors [40], Haar-HOG cascade [2], and Blob-HOG cascade methods are proposed. Due to the application of background subtraction, these methodologies are not able to recognize idle resources. Park and Brilakis [36] also proposed HOG and HSV color histograms together with background subtraction for initializing vision trackers for workers.

In the computer vision community, there is a large number of emerging works in the area of object detection and human pose estimation [9,10,14,27,43]. The results of these algorithms seem to be both effective and accurate. For example, Felzenszwalb et al. proposed method [14] that can detect objects with deformable configurations, which can be very effective for action recognition purposes. Moreover, this algorithm is able to detect different parts of objects and has potential for detecting occluded resources in site video streams. The work proposed in van de Weijer and Schmid [41] extended the description of local features with color information. The results of this study show that color descriptors remain reliable under certain photometric and geometrical changes, and with decreasing image quality. Although existing computer vision methods show very promising results, in most cases they are only applied and validated under controlled settings. We have also extensively tested their direct applications and in the most cases where occluded and dynamic video streams were used, an acceptable precision level for construction performance assessment purposes was not obtained. Nevertheless, certain elements of these works can be effectively used to create new techniques for automated worker and equipment detection and tracking.

There is a need for techniques that can support automated 2D detection and localization of construction workers and equipment even when they are *idle*. This enables development of both action recognition and 3D tracking methods, which can ultimately bring awareness on project specific issues, empower practitioners to take corrective actions, avoid delays, and minimize excessive impacts due to low operational efficiency or unsafe practices.

In this paper, we built upon our previous work on application of HOG detectors in intensity-based video frames [30] and propose a new method based on multi-scale detection window and joint representation of HOG and color values in the Hue–Saturation-Value (HSV) space, which we denote as HOG + C. The details of our proposed method are outlined in the following.

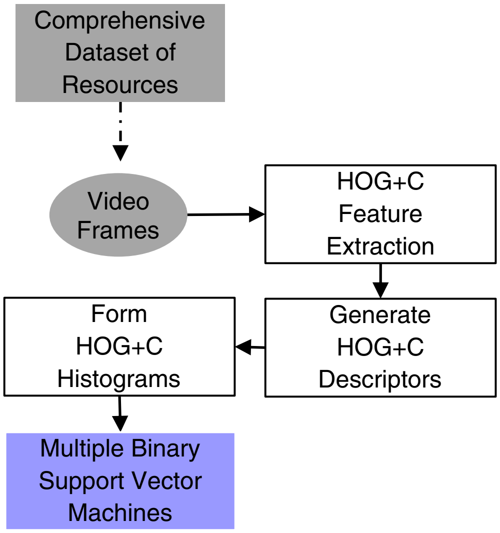
Comprehensive Dataset of Resources
Video Frames
HOG+C Feature Extraction
Generate HOG+C Descriptors
Form HOG+C Histograms
Multiple Binary Support Vector Machines

Fig. 2. The flowchart of the proposed method.

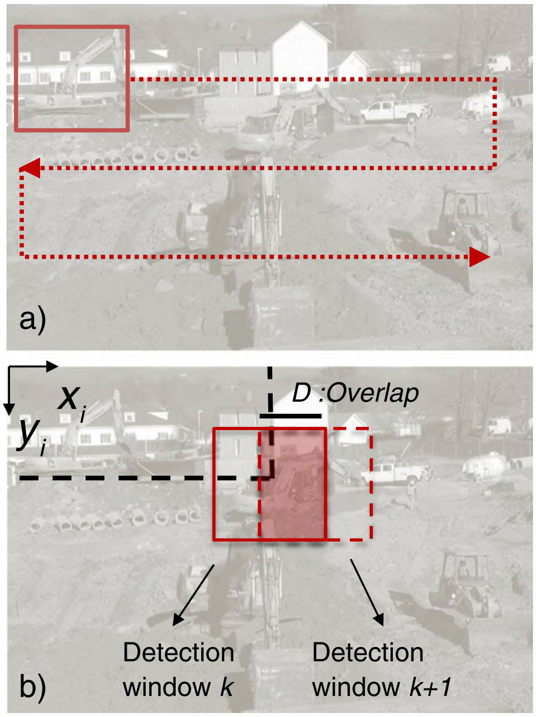

Fig. 3. Representation of sliding detection window.

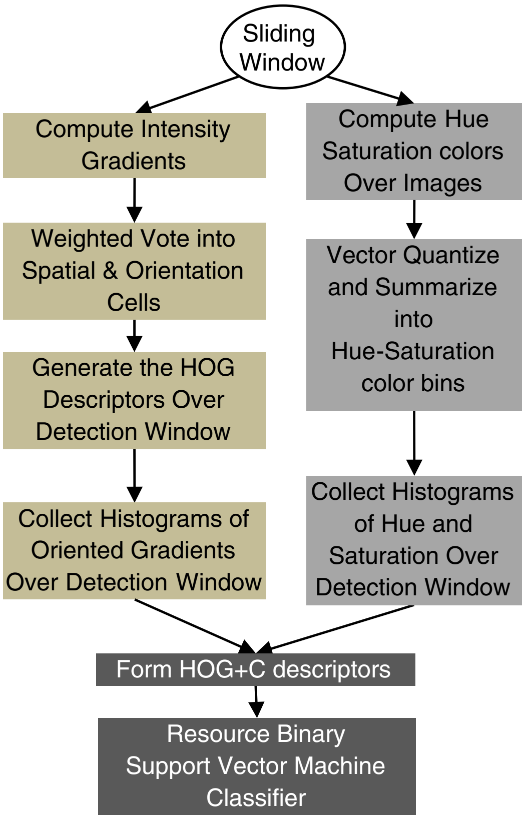
Sliding Window
Compute Intensity Gradients
Weighted Vote into Spatial & Orientation Cells
Generate the HOG Descriptors Over Detection Window
Collect Histograms of Oriented Gradients Over Detection Window
Compute Hue Saturation colors Over Images
Vector Quantize and Summarize into Hue-Saturation color bins
Collect Histograms of Hue and Saturation Over Detection Window
Form HOG+C descriptors
Resource Binary Support Vector Machine Classifier

Fig. 4. The algorithm for automated detection of construction resources.

3. Overview of the proposed method

Given 2D video frames collected with fixed cameras on construction sites, our goal is to 1) automatically learn visual classifiers for different equipment and workers and 2) apply the learned models to perform detection and classification of equipment and workers in new video frames. The proposed approach is illustrated in Fig. 2.

It is assumed that the video frames contain typical dynamic construction foregrounds and backgrounds that can generate occlusions. The training stage in our work is supervised in the sense that we annotate bounding boxes around each equipment/worker in the image. During the testing stage, the proposed method automatically places the bounding boxes and can handle observations containing more than single resource under various degrees of occlusion.

Large variations in illumination, weather conditions, and resolution, in addition to the scale of workers and equipment in 2D video streams and their intra-class variability (particularly in the case of equipment) makes site video streams challenging to work with. In order to address this problem, we introduce 1) multi-scale sliding detection windows, and 2) HOG + C descriptors which are formed by concatenating HOG [9] with Histograms Of Color (HOC) to create an automated 2D detection method. These steps are described in the following subsections.

3.1. Multi-scale sliding detection window

Our method for detection of workers and equipment involves application of a *sliding* detection window. The basic idea is that the detection window scans across each video frame at all positions and for several spatial scales to find the best candidates. As shown in Fig. 3, during this process each window is independently analyzed and classified whether it contains a particular type of resource or not. This strategy provides two key benefits:

1) *Detection of workers and equipment while idle*; i.e., it examines static windows for possible resource candidates and is not limited to the detection of moving foreground objects (typical in background subtraction techniques);
2) *Detection of workers and equipment in close proximity of each other under high degrees of occlusion*; several overlapping windows can be chosen as the best candidates for construction resources which is a key component required for safety assessments.

In the following, the process of detecting workers and equipment within each detection window is described.

3.2. Resource detection and classification for each detection window

Fig. 4 presents an overview of our method for learning and detection of workers and equipment within each candidate detection window. As observed in the figure, we extract two types of visual information: 1) image gradients via a HOG descriptor (left side of Fig. 4); and 2) color cues captured by a HOC descriptor (right side of Fig. 4). Once these descriptors are formed, they are combined and fed into a machine learning classifier to identify whether or not the detection window contains a resource of interest.

3.2.1. Histogram of oriented gradients (HOG)

The main idea is that the local shape and appearance of workers and equipment in a given detection window can be characterized by distribution of local intensity gradients. These properties can be captured via HOG descriptors [9]. In order to do so, we first compute the magnitude $|\nabla f(x,y)|$ and orientation (angle) $\theta(x,y)$ of the intensity gradient $|\nabla f(x,y)|$ for each pixel within the detection window. Next, we vector quantize and summarize all these orientations and their magnitudes within the detection window into a HOG. More precisely,

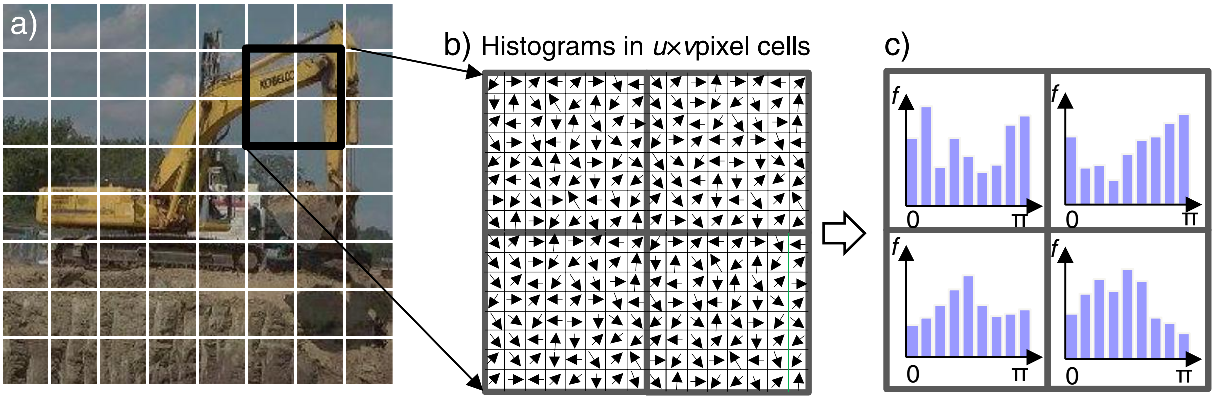

Fig. 5. The formation of the Histogram of Oriented Gradients for each detection window: (a) a 250×250 pixel detection window, (b) 4×4 pixel cells in each window, and (c) the Histogram of Oriented Gradients corresponding to 4 cells.

```
Data:  Training sets $(x_1,l_1),\ldots,(x_n,l_n)$ where $l_i = 0,1$ for
         negative and positive examples of each
         resource respectively
Results: a one-vs.-all SVM Model ($M$) for each resource
for each positive and negative example file list
  for each cell in an example
      create HOG descriptor
      create Hue-Saturation color descriptor
      concatenate into HOG+C descriptor
  end for
  append to the train_positive and train_negative lists
end for
find support vectors
return $M$
```

Fig. 6. Algorithm for training process.

the detection window (Fig. 5a) is divided into $t_x \times t_y$ local spatial regions (*cells*) where each cell contains $u \times v$ pixels (Fig. 5b). Each pixel casts a weighted vote for an edge orientation histogram bin, based on the orientation of the image gradient at that pixel. These votes are then accumulated into n evenly-spaced orientation bins over the cells; i.e., each bin characterizing an unsigned gradient: $i \times \pi/n$ $i = 1,\ldots,n$ (see Fig. 5c). A naïve distribution scheme in form of voting to the nearest orientation bins creates aliasing effects due to under-sampling. Similarly, pixels near the cell boundaries can also produce aliasing along spatial dimensions. To reduce aliasing, similar to [8], the gradient magnitudes at the pixel level are interpolated bilinearly between the neighboring bin centers in both orientation and position. The outcome of this process is a HOG descriptor for each detection window. Inspired by [14], we use an augmented low-dimensional HOG feature set that includes both contrast sensitive and insensitive features, leading to a 31-dimensional feature vector. By comparing these low-dimensional feature vectors with their original 36-dimensions introduced in Dalal and Triggs [9], Felzenszwalb et al. [14] showed that the performance of the HOG descriptors could be improved; which is the rationale behind their application in our method.

3.2.2. Histogram of hue–saturation colors (HOC)

Simultaneous to the formation of the HOG descriptor, a histogram of colors (HOC) is also generated. In order to maintain invariance to illumination changes and inspired by van de Weijer and Schmid [41], instead of using Red–Green–Blue (RGB) color space, in our algorithm, we use Hue–Saturation-Value (HSV) colors [45]. It is hypothesized that using hue and saturation components instead of RGB can improve the detection of construction workers and equipment in saturated construction scenes (this hypothesis is validated in Section 5 of this paper). After converting the image into the HSV space, we only keep the hue and saturation components, which are summarized by a histogram that counts the occurrences of a set of evenly spaced normalized hue and saturation values. In all our experiments, we vector-quantize the color space into 6 bins for hue and 6 bins for saturation to generate HOC descriptors which is in form of a 2D histogram with 36 bins. These descriptors over the detector window are locally historgrammed and concatenated with the HOG to form the HOG + C descriptors.

```
Data:  Learned one-against-all SVM Model ($M_j$) per resource
        Testing images and video frames
Results: bounding boxes and classification scores per box
for each image in the testing list
   for each spatial scale in search scale
      resample image to new scale
      for each detection window in an image
         compute HOG descriptor
         compute Hue-Saturation color descriptor
         concatenate and from the HOG+C descriptor
         analyze the detection window with $M_j$
      end for
   end for
   non-maximum suppression
end for
return bounding boxes and classification scores
```

Fig. 7. Algorithm for detection (testing) process.

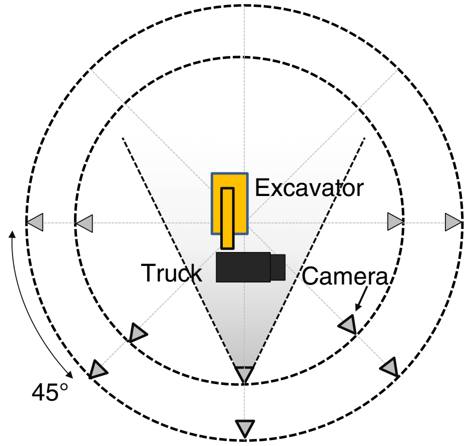

Fig. 8. Data collection and experimental setup.

3.3. Support vector machine (SVM) classifier

Once the HOG + C descriptors are formed, they are placed into a machine learning classifier to identify whether or not the detection window contains a given resource. For this purpose, we use multi-class Support Vector Machine (SVM) classification approach [7]. Given n labeled training datapoints $\{x_i, y_i\}$, wherein x_i $(i = 1,\ldots,n,\ x_i \in R^d)$ is the set of d-dimensional HOG + C descriptors computed from each image example (i), and $y_i \in \{0,1\}$ is the binary class label (e.g., worker or non-worker), the SVM classifier aims at finding an optimal hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ between the positive and negative samples. We assume that there is no a priori knowledge about the distribution of the resource class video frames. Hence, the optimal hyper plane is one that maximizes the geometric margin γ as follows:

$$\gamma = \frac{2}{||w||}. \tag{1}$$

For each binary SVM resource classification, the dataset contains considerable number of video frame entries. Hence the training data will be linearly separated and as a result the classification can be formulated as:

$$\begin{gathered}\min_{w,b}\frac{1}{2}||w||^2 \\ \text{subject to}:\ y_i(w.x_i + b) \geq 1 \text{ for } i = 1,\ldots N.\end{gathered} \tag{2}$$

The presence of noise and occlusions which is typical in construction site video streams produces outliers in the SVM classifiers. Hence the

Table 1
The number of positive and negative image samples used for training and testing construction resource classifiers.

Resource	Dataset	Positive	Negative
Excavator	Training	1895	2280
	Testing	1008	746
Truck	Training	1212	2434
	Testing	738	1122
Worker	Training	1840	2487
	Testing	702	1043

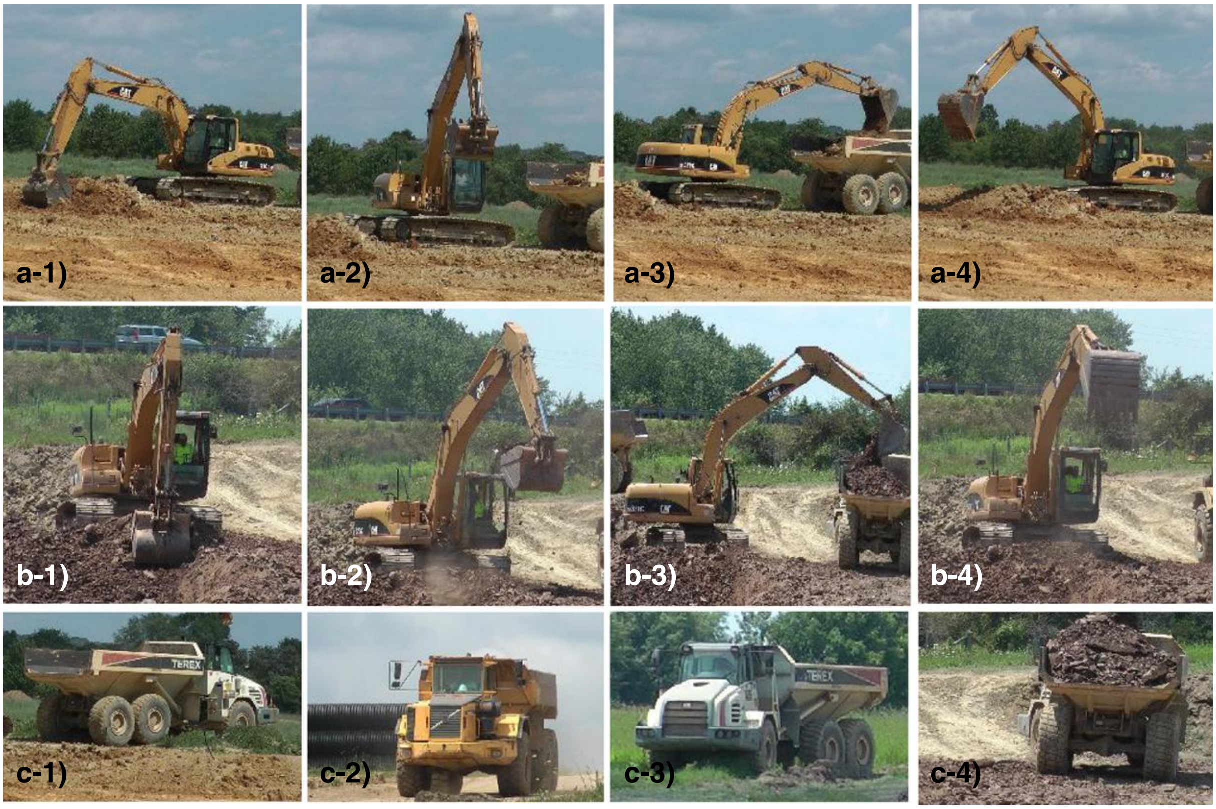

Fig. 9. Example frames from video sequences of excavator and truck operations. From left to right in rows (a) and (b): digging, hauling, dumping, and swinging action classes which illustrate tremendous appearance changes because of variability in equipment part arrangement. Row (c) shows the appearance changes due to view point location for a truck (e.g., side view, frontal view).

slack variables ξ_i are introduced and consequently the SVM optimization problem can be written as:

$$\begin{aligned} &\min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to: } &y_i(w.x_i + b) \geq 1-\xi_i \text{ for } i = 1,\ldots,N \\ &\xi_i \geq 0 \text{ for } i = 1,\ldots,N. \end{aligned} \tag{3}$$

In this formula, C represents a penalty constant that can be determined by a cross-validation technique. As observed in Fig. 6, the inputs to the learning (training) algorithm are the training examples for different resources and the outputs are the trained models for detection of various resources.

To effectively classify the testing images with the HOG + C descriptors, it is necessary to slide the detection window over each image at multiple spatial scales. This is accomplished by rescaling the image and enabling the detection window to search at different scales. For each spatial scale of the detection window, image gradients and hue–saturation components are calculated, and the resulting feature vector is classified using the learned one-against-all SVM model. If the classification is positive, the bounding box for the detection window and the classification value (i.e., classifier score) are added to a list for further processing. Next, the detection window is moved across the entire video frame using a specified search step; i.e., m pixels. In this paper, these spatial steps are referred as the detection window overlaps. Once all detection window positions have been classified for all spatial scales, the positively detected bounding boxes are processed using a non-maximum suppression technique. The width of the bounding boxes and the distance between box centers are used to determine if an adjacent bounding box needs to be considered as a neighbor for non-maximum suppression. The final outcome of this step is a set of bounding boxes which capture all positive classifications and their scores. Fig. 7 shows the algorithm for the detection (testing) process of a resource classifier.

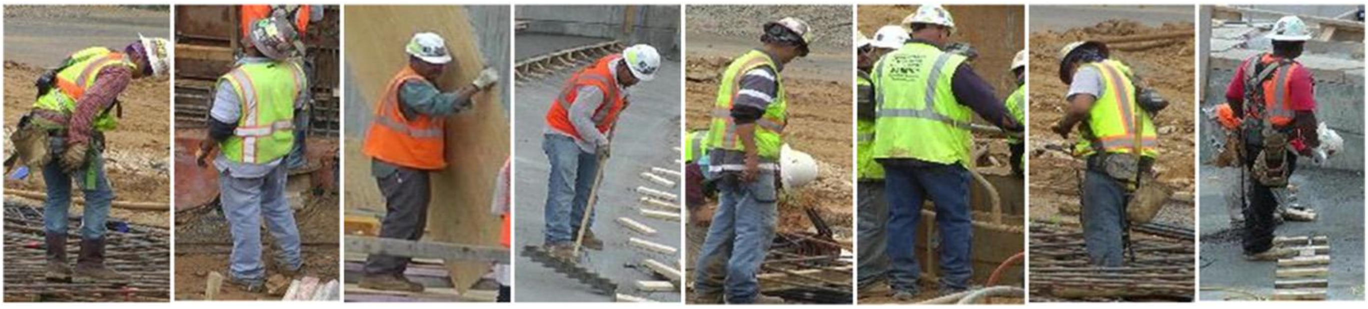

Fig. 10. Example frames of various pose from worker operation class. These examples exhibit appearance changes due to body part arrangements and self-occlusions among body parts (e.g., one hand fully occluded in the first and third frames from left).

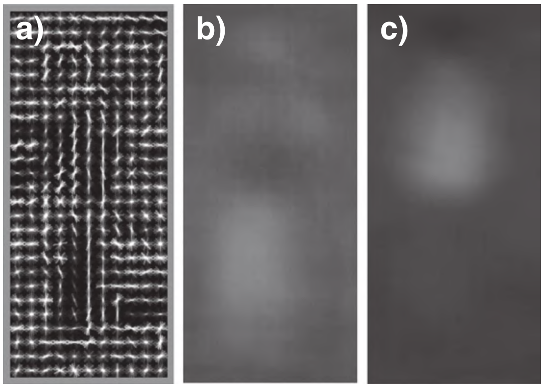

Fig. 11. (a–c) The average oriented gradients, average hue values, and average saturation values over the worker dataset.

4. Experiments and validation metrics

4.1. Data collection and experimental setup

Due to the lack of existing datasets for benchmarking visual detection of construction workers and equipment, it was necessary to create a new comprehensive database. This dataset is for both training and testing purposes so that it can be released to the community for further development and validation of new algorithms. For this purpose, we collected 300 h of video streams that were recorded from five different construction projects (i.e., two building and three infrastructure projects). In order to create a comprehensive dataset with varying degrees of viewpoint, scale, and illumination changes, the videos were collected over the span of six months. Due to various possible appearances of equipment, particularly, their actions from different views and scales in a video frame, as shown in Fig. 8, several cameras were set up in two 180° semi-circles (each camera roughly 45° apart from one another). This strategy enables the resources to be videotaped at two different scales (full and half high definition video frame heights). Combined with the strategy used to encode spatial scale in the sliding detection window, all possible scales are considered.

Our equipment dataset contains three types of excavators (manufacturers: Caterpillar, Komatsu, and Kobelco) and three types of dump trucks (manufacturers: Caterpillar, Trex, and Volvo). In the case of workers, the dataset was collected from concrete placement and steel erection operations and mainly contains *standing* crews. Given the significant difference in body configuration of bending workers, we assumed that those need to be trained separately and hence were left out from the scope of this paper.

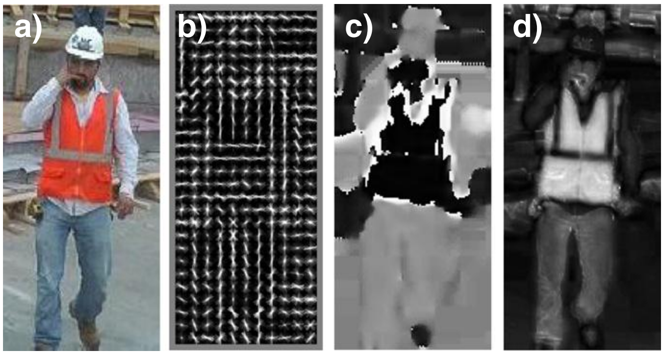

Fig. 13. Example of testing for worker dataset: (a) a test image, (b) the oriented gradients, (c) hue map and (d) saturation map.

Table 1 shows the size of the training and testing datasets. As observed, a total of 2903, 1952, and 2653 positive High Definition (HD) frames (frames that represent an actual resource class) were manually segmented, labeled, and used for initial experiments on excavators, trucks, and workers respectively. These frames were randomly divided into two groups of training and testing by a ratio of 2 to 1. Training frames is cropped to contain only single resources, however in testing phase there is no such a constraint and frames can contain multiple resources. The negative images for each binary classification include: (a) the positive instances from the other two classes and (b) additional 500 negative frames that represent typical dynamic backgrounds from construction sites and may include other resources. Positive frames refer to those frames that contain the object of interest, while negative frames refer to frames that do not contain the object of interest. The classifiers for each resource were individually trained using their corresponding training datasets and were evaluated using the testing dataset. The entire dataset is made public at: http://www.raamac.cee.vt.edu/detectiontracking.

4.2. Performance evaluation measures

To quantify and benchmark the performance of the 2D detection algorithm, we plot the Precision–Recall and Detection Error Tradeoff (DET) curves. DET curves illustrate the relationship between miss rates versus FPPW (False Positive Per Window) and are introduced by National Institute of Standards and Technology (NIST) [29]. Both of these

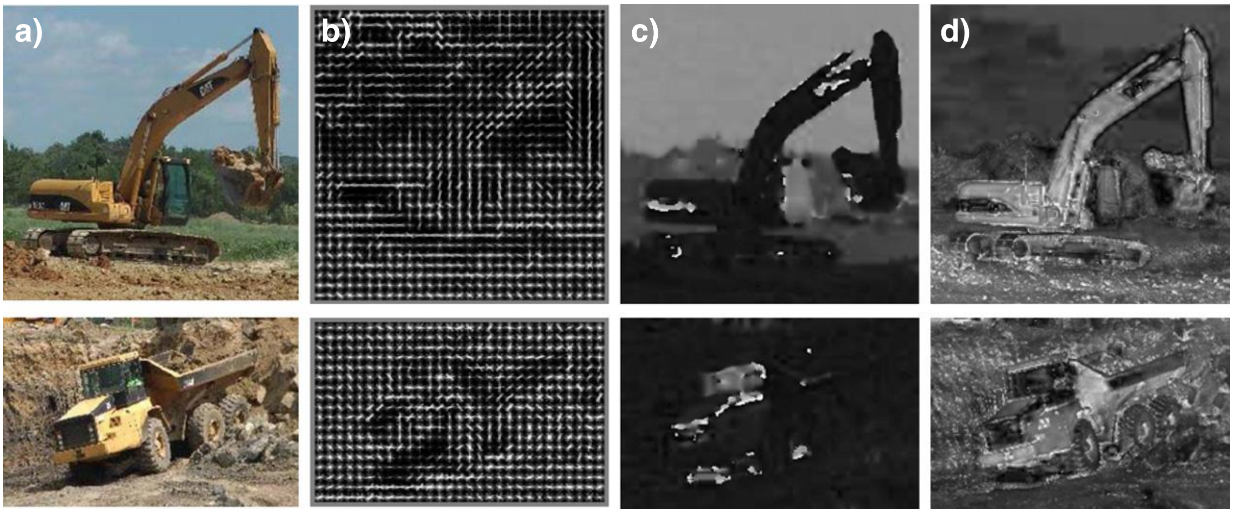

Fig. 12. Example of testing for excavators and trucks datasets, respectively: each row: (a) a test image, (b) the oriented gradients, (c) hue map and (d) saturation map.

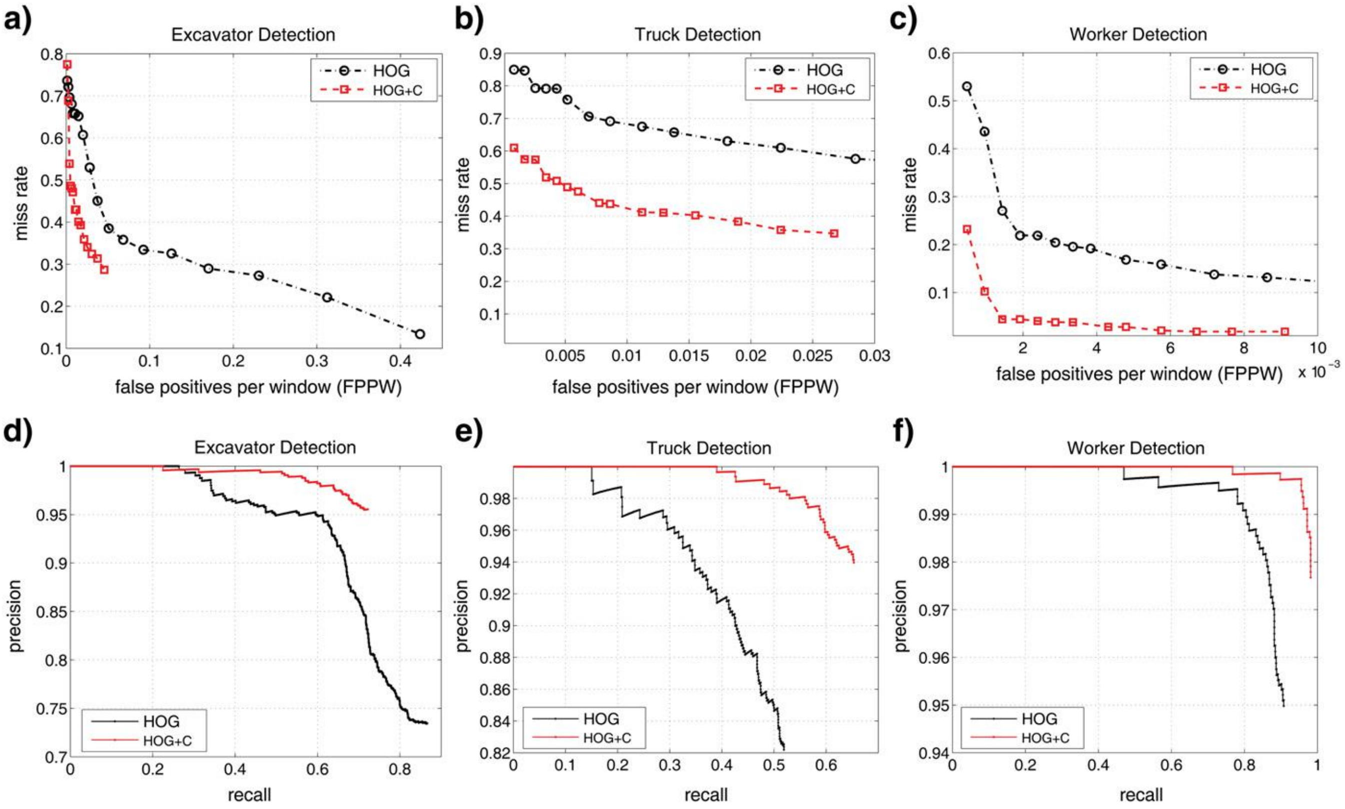

Fig. 14. Overall results on performance of HOG and HOG + C on detection of construction resources. (a–c) DET and (d–f) precision–recall curves for detection of excavators, trucks, and workers, respectively.

evaluation metrics are extensively used in the Computer Vision community. In particular, methods that use the sliding detection window technique for pedestrian detection commonly use DET curves for evaluation. These metrics are both set-based measures; i.e., they evaluate the quality of an unordered set of data entries. In the context of 2D detection of construction resources, we define each as follows.

4.3. Precision–recall curve

To facilitate comparing the overall average performance of the variations of the proposed 2D tracking algorithm over a particular set of video frames, individual detection class *precision* values are interpolated to a set of standard *recall* levels (0 to 1 in increments of 0.1). Here, precision is the fraction of retrieved instances that are relevant to the particular classification, while recall is the fraction of relevant instances that are retrieved. Thus, precision and recall are calculated as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

where in TP is the number of True Positives, FN is the number of False Negatives and FP is the number of False Positives. For instance, if the *worker* detection window recognizes a worker, it will be a TP; if an equipment instance is incorrectly recognized under worker class, it will be a FP. When a worker instance is not recognized under the worker class, then the instance is a FN. The particular rule used to interpolate precision at recall level *i* is to use the maximum precision obtained

Table 2
Average accuracies for detection of different construction resources (%).

Resources	HOG	HOG + C
Worker	96.07	98.83
Excavator	74.28	82.10
Truck	76.92	84.88

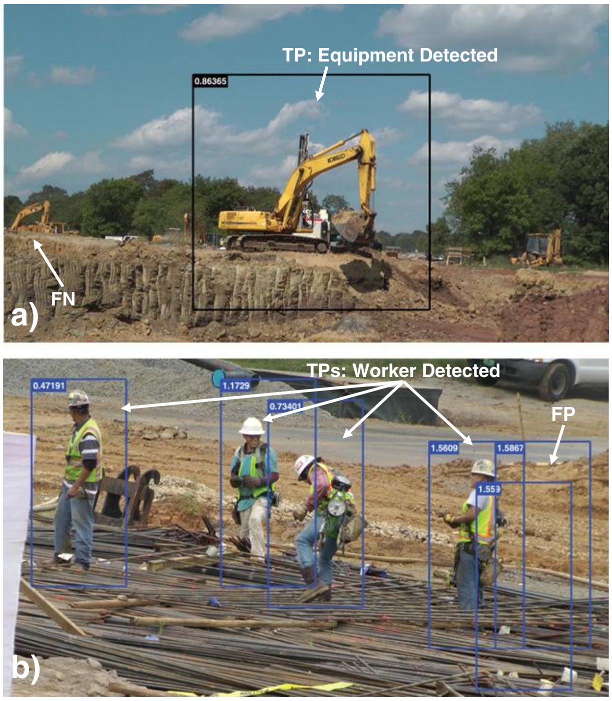

Fig. 15. Example of TP, FP, and FN in detection of construction resources (the top left boxes show the classification scores).

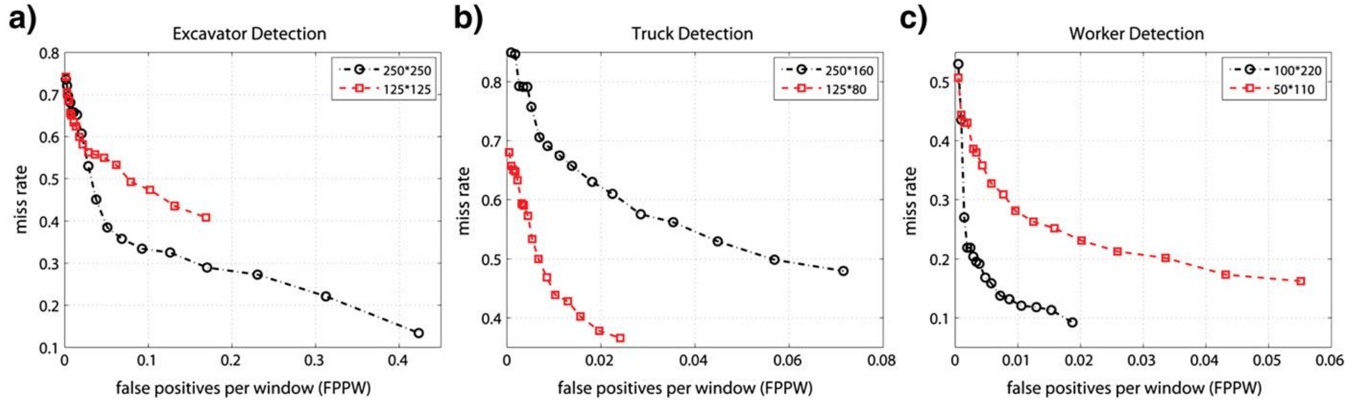

Fig. 16. Effect of the detector window size on performance of HOG for detection of different construction resources.

from the detection class for any recall level great than or equal to *i*. For each recall level, the precision is calculated; then the values are connected and plotted to form a curve.

4.4. Detection error tradeoff curve

For sliding detection window techniques, the DET curves allow the performance of the algorithms to be compared more easily. Based on these curves, a better performance of the detector should achieve minimum *miss rate* and *FPPW* (the curve will be closer to the lower-left corner). The terms *miss rate* and *FPPW* are defined as follows:

$$\text{miss rate} = 1 - \text{recall rate} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{FPPW} = \frac{\text{FP}}{\text{TN} + \text{FP}}. \tag{7}$$

When necessary, the average accuracy of the resource detection is also calculated using the following formula:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{8}$$

5. Experimental results

In this following section, we first present the experimental results from our proposed algorithm. In the subsequent subsections, we test the efficiency of our approach on various model parameters. As observed in Figs. 9 and 10, our database includes video frames from multiple resources. Each frame shows a different body configuration and is captured from a unique scale under a specific pose, illumination and occlusion condition.

We implemented the proposed algorithms in MATLAB with several components in C++ for faster computation. The performance of our implementation was benchmarked on a Linux 64 bit platform with 24 GB memory and 3.2 GHz Core i7 CPU. In our proposed method, the detectors have the following properties:

- The sizes of the detection windows for excavators, trucks, and workers are set to 250×250, 250×160, and 100×220 pixels respectively;
- Linear gradient [−1;0;1] voting into 9 orientation bins in 0–180° is used for generating all HOG descriptors; i.e., visually symmetrical gradients are chosen for detection of construction resources;
- L2-normalized blocks with 4 cells containing 8×8, 4×4 and 16×16 pixels were used to generate HOG descriptors for excavators, dump trucks, and workers respectively; and finally,
- Linear SVM classifiers with $C = 1$ are used for the detection and classification of each resource.
- The time required for testing on the HD image is around 10 min.

Fig. 11 shows a HOG + C descriptor which is learned using the worker training dataset. Figs. 12 and 13, each show an example of a testing image, in addition to their HOG + C descriptors.

In our testing phase, the detection window slides at multiple uniform scales (i.e., 1.0, 2.0, and 3.0×). This strategy not only allows resources with smaller scales to be detected, but also enables the method to be used on lower quality site video streams. Fig. 14 shows the DET and precision-recall curves for both HOG + C and HOG detectors and compares their performances for all three categories of resources on testing dataset. As observed, the new method based on HOG + C descriptors significantly improves the performance of detecting construction resources.

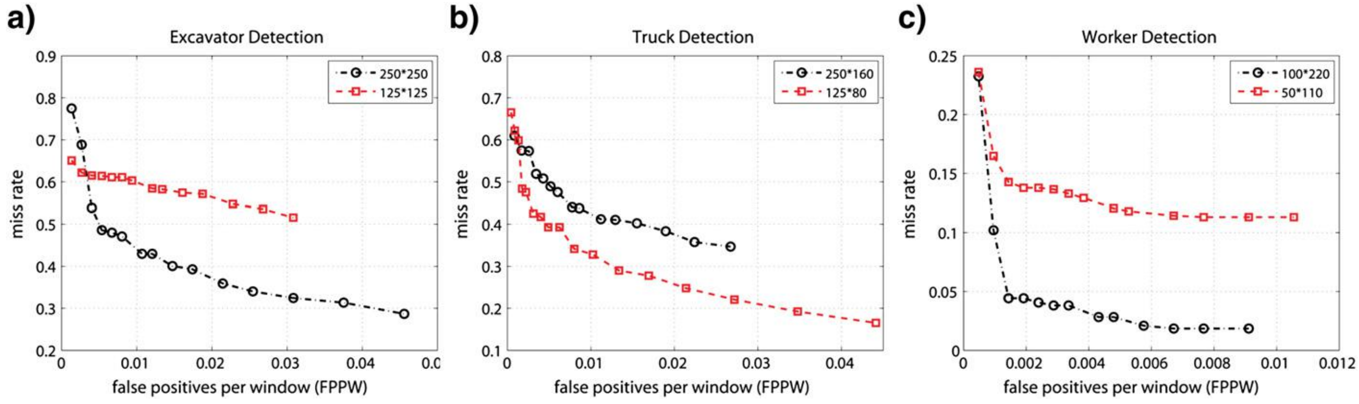

Fig. 17. Effect of the detector window size on performance of HOG + C for detection of different construction resources.

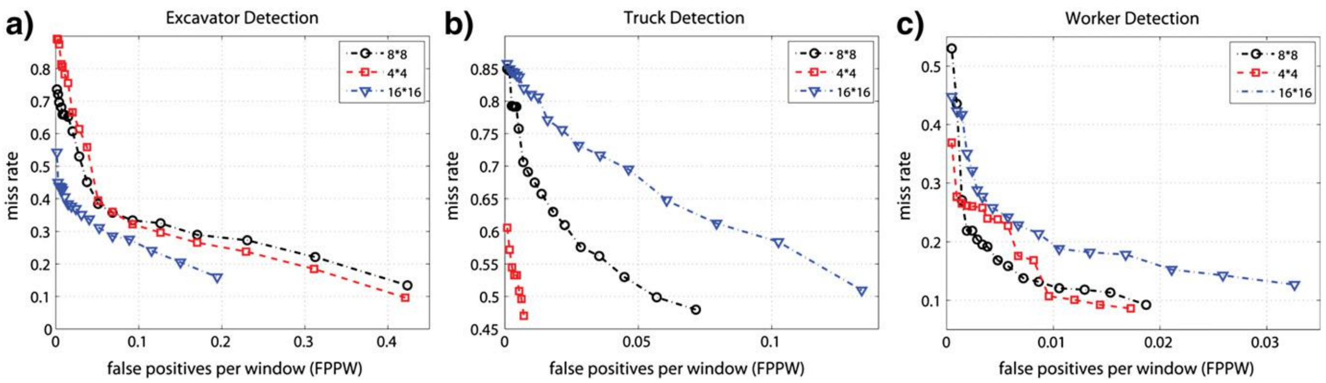

Fig. 18. Effect of the cell size on performance of HOG for detection of different construction resources.

In particular it achieves lower miss rates in lower FPPWs and also higher precisions in higher recall values.

The average accuracies in detection of each resource are listed in Table 2. Using HOG for the detection of workers has a higher average accuracy compared to the excavators and trucks. This is due to the consistent pose of the standing workers in the worker dataset compared to the excavators and trucks. In our method, we have view-independent models for excavators and trucks; i.e., all possible viewpoints are considered together. As a result, our HOG-only classifiers result in lower accuracies. Nevertheless, due to the distinct colors of equipment, adding the color information and forming HOG + C histograms significantly improves their performance.

Several examples of TP, FP, and FN for different resources detection methods are presented in Fig. 15. As seen in the Fig. 15a, the detected excavator is labeled as a TP. In this video, at far end left, a half occluded excavator is observed. Due to the small scale in the video frame, this excavator is not detected by our algorithm and is labeled as a FN accordingly. Fig. 15b shows an example from the worker detection process. Here, four workers were accurately detected (TPs). A false alarm (FP) is also observed wherein the background is detected as the worker.

5.1. Discussion on model parameters

In the following subsections, we systematically study the effects of the various choices on both HOG and HOG + C detectors. Particularly the effect of the size of the detection window and cells, and the number of bins in HOC descriptors are studied in detail. The effect of using various percentages of overlaps for the detection windows is also further explored. The best parameters from these experiments which were presented in Section 4 were selected based on the highest average performances and most reasonable computational times.

5.1.1. Effect of the sliding detection window size

Figs. 16 and 17 show the effect of detection window size on the performance of HOG and HOG + C descriptors for excavator, truck, and worker classes respectively. In the case of detecting dump trucks (see Figs. 16b, 17b), 250 × 160 and 125 × 80 pixel detection windows were used to evaluate the performance. As observed, smaller windows perform better in the detection of the dump trucks, while the performance degrades in the case of workers and excavators (see Figs. 16a, c, 17a, c). In the case of workers and excavators, a large window size is needed to statistically capture the changes of intensity for different postures within various actions. However in the case of dump trucks, the actions are more simple, and hence a smaller window can better capture the changes of intensity. Overall, smaller size detectors, enable the method to detect those resources that are far from the video camera and/or appear in low-quality video streams.

5.1.2. Effect of cell size on the detection performance

Another effective factor on the performance of our resource detector is the size of the cells. We evaluated three different sizes for the cells: 4 × 4, 8 × 8, and 16 × 16 pixels. Figs. 18 and 19 demonstrate and compare the performance of HOG and our HOG + C detectors with varying cell sizes. As observed in Figs. 18b and 19b in the case of detecting dump trucks, the 4 × 4 cell resulted in the best performance. While in the case of workers and excavators, the 8 × 8, and 16 × 16 cells performed better respectively. Among our resource categories, the detection of dump trucks is more challenging. Due is to the notion that their appearance significantly differs from one truck to another. Since the pose of the truck can also have a significant impact on their 2D visual appearance, their detection using view-independent HOG + C descriptors, in particular in clutter backgrounds is more challenging.

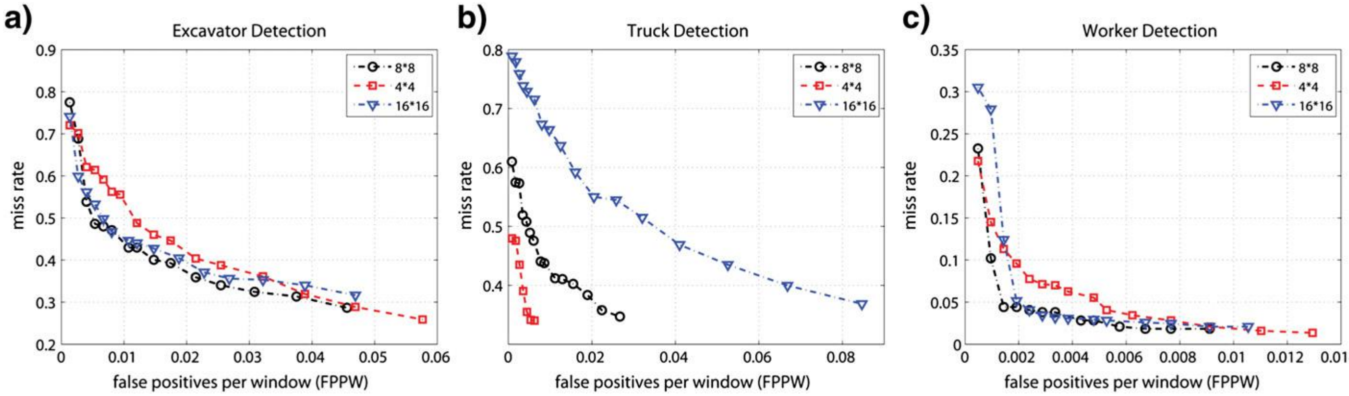

Fig. 19. Effect of the cell size on performance of HOG + C for detection of different construction resources.

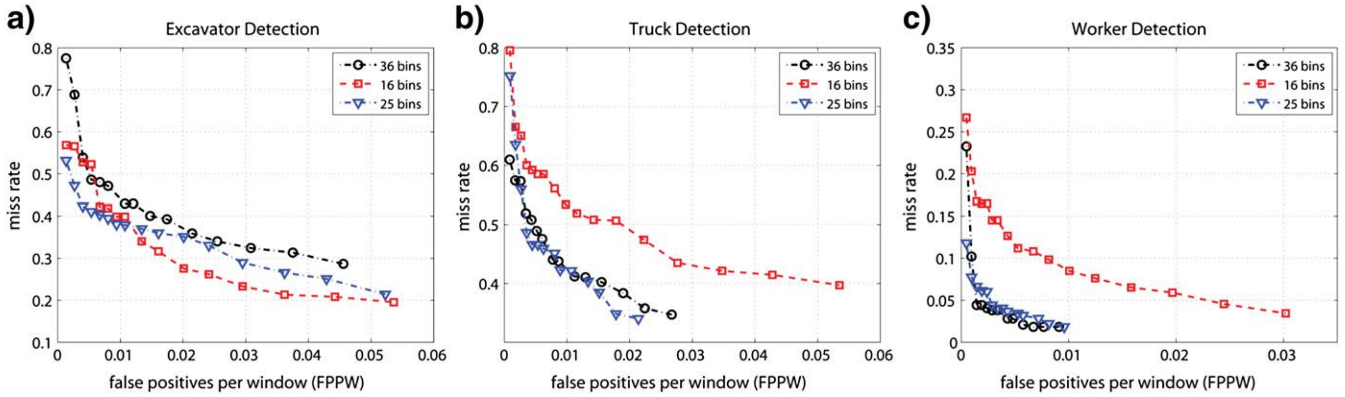

Fig. 20. Effect of the number of bins in HOC on performance of HOG + C for detecting different construction resources.

5.1.3. Effect of number of bins in HOC on detection performance

Finally, we evaluated the effect of the number of bins in HOC descriptors to find out which combination results in the best detection performance. Fig. 20 demonstrates the outcome of this comparative study. In particular, the effect of three different numbers of bins (16, 25, and 36) was studied. As observed, the 25 and 36 bin HOC descriptors outperform others in the detection of dump trucks and workers respectively. In the case of excavators the 16 bin HOC descriptors showed the best performance.

5.2. Resource detection using the sliding detection window

In the previous section, we evaluated the performance of our detection method on isolated video frames in which the expectation was to detect a single resource. Here we focus on evaluation of our method for detection of multiple resources with varying degrees of occlusion. The ability of analyzing multiple overlapping windows in our method 1) increases the accuracy of 2D localization, and 2) enables detection of multiple resources in close proximity to one another, all in a reasonable computational time.

Fig. 21 shows the impact of different level of window overlap on accuracy of detecting an excavator in noisy construction backgrounds. As illustrated, increasing the percentage of overlap between detection windows from 0% (without overlap) to 98%, significantly improves the accuracy of localizing resources in 2D. Obviously growing the percentage of overlap between detection windows increases the number of FPs. In order to achieve reasonable performance we used a non-maximum suppression step to select only those detection windows that are returning the highest scores. We also performed a trade-off analysis between the percentage of window overlap, accuracy of 2D localization, and computation time. On large images containing multiple resources, an overlap of 90% resulted in the most reasonable 2D localization accuracy considering the computation time.

One of the key challenges in automated tracking of construction resources is the ability to continuously detect the resource in video frames wherein the equipment pose, illumination and occlusion are rapidly changing. Fig. 22 shows the performance of our algorithm for detection of an excavator in a video sequence where the pose of the equipment was rapidly changing.

Figs. 23, 24, and 25, show the performance of our detector window in detecting multiple resources. As illustrated in Fig. 23, the 90% overlap, enables excavators and trucks that are working in close proximity to each other to be robustly detected. Fig. 24 shows another example on detection of construction crew working in proximity to an excavator. This is a critical component for safety assessment purposes. Fig. 25 shows the performance of our algorithm in detection of multiple excavators in different distance from the camera (scale) and from multiple viewpoints. Detecting multiple '*parts*' for resources and using those as indicators for tracking under severe occlusions is under study.

6. Discussion on the proposed method and research challenges

This study presented the first comprehensive video frame dataset for 2D detection of excavators, dump trucks, and standing construction workers. The average accuracies of the detection obtained for workers, excavators, and dump trucks are 98.83%, 82.10%, and 84.88% respectively. The ability to detect idling resources, distinguishes our work from

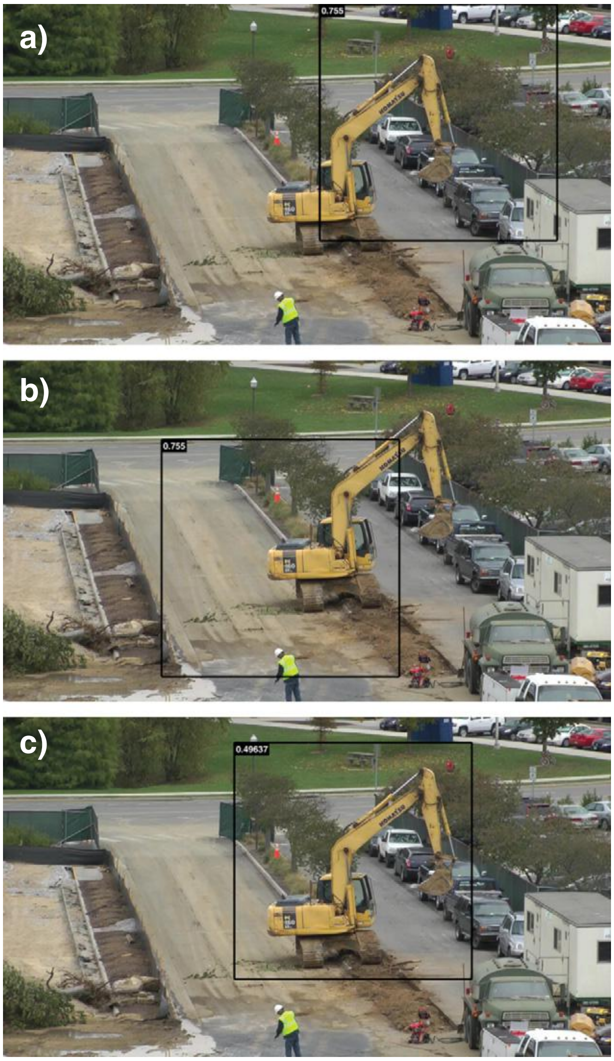

Fig. 21. Effect of the detection window overlap in accuracy of localizing construction resources in 2D: (a) without overlap, (b) 50% overlap and (c) 98% overlap.

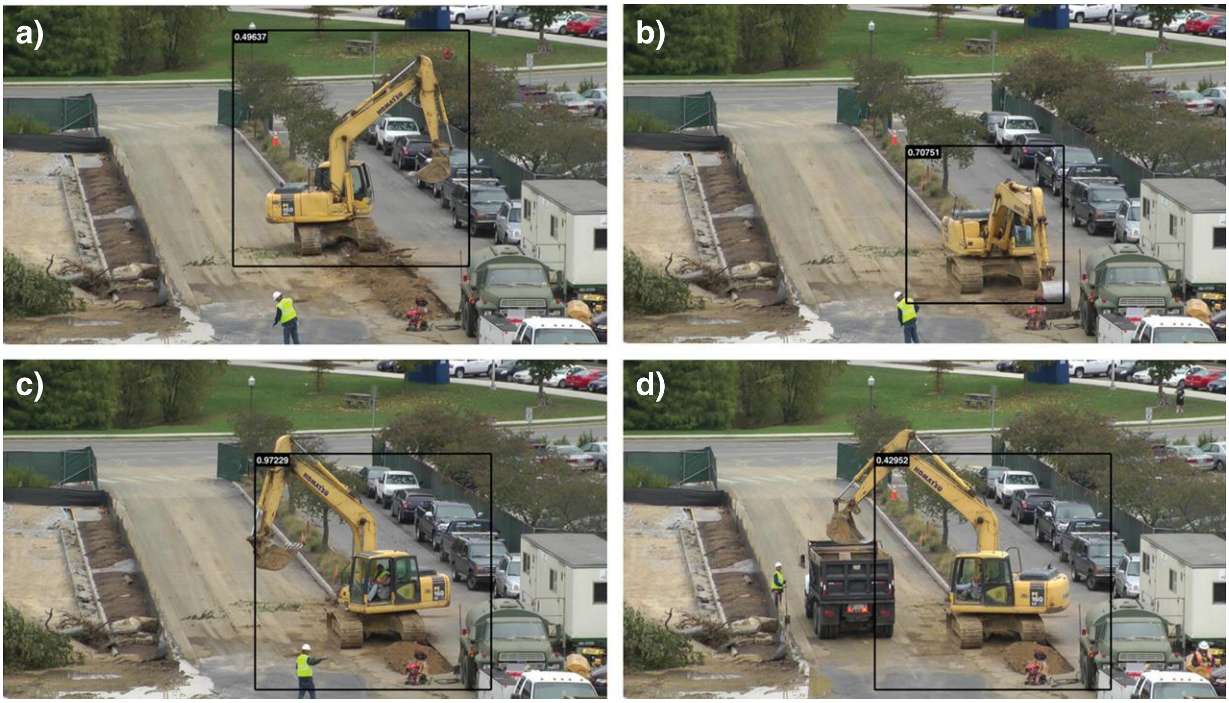

Fig. 22. Detecting an excavator in a video sequence wherein the pose is rapidly changing.

previous methods presented in the AEC community. The results also indicate the robustness of the method to dynamic changes of illumination, viewpoint, camera resolution, and scale. It further shows reasonable robustness to static and dynamic occlusions. The minimal detectable spatial resolution of the equipment in videos in the range of (80–800)× (80–800) and (100–800)×(100–800) pixels per excavator and dump truck, and (50–700)×(50–700) pixels per worker, promises the applicability of the proposed method for existing site video cameras. While this paper presented the initial steps towards processing site video streams for the purpose of 2D resource detection and localization, several critical challenges remain. Some of the open research problems for our community include:

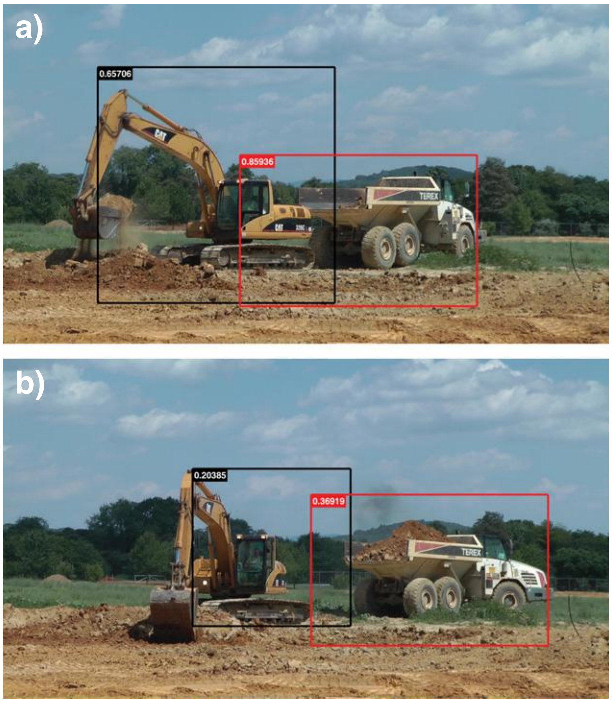

Fig. 23. Detection in a video sequence where in an excavator and a truck are working in the proximity of each other.

- Real-time 2D detection and localization in long video sequences. The presented algorithm is capable of accurately tracking resources in a post processing stage, which makes it attractive for development of action recognition methods. Nonetheless, for safety analysis, there is a need for real-time 2D detection and localization. The current high computation time in our method is inherent to the application of sliding detection windows which were primarily created to handle detection of idling resources. To detect and track construction resources in real-time, more work is needed to implement the HOG+C based sliding window algorithm using the NVIDIA CUDA parallel computation framework.
- Equipment detection and localization over a network of fixed cameras. 3D tracking for multiple resources requires *precise* 2D detection and localization in each video camera and subsequent matching across all views. Given the distance of the cameras to resources on the jobsite, small deviations in 2D localization can generate large error in 3D localization. There is a need for methods that can identify several parts or features within the detection windows across all video cameras to enable high precision triangulation in 3D. Detecting geometrically and visually consistent correspondences across multiple cameras can also form several hypotheses for each detection and enable development of algorithms that can choose best hypothesis for classification. It further minimizes the effect of noise caused by lateral movement of the camera, and the dynamic motions of foreground or background.
- Variability in equipment type/models and worker body postures. Highly accurate 2D detection requires comprehensive datasets of all type/models of equipment and various worker body postures to be collected for training purposes. The dataset presented in this work only includes two types of equipment from six different manufacturers, and standing workers. Development of larger datasets for equipment and workers with different body postures (e.g., bending, sitting) is needed.

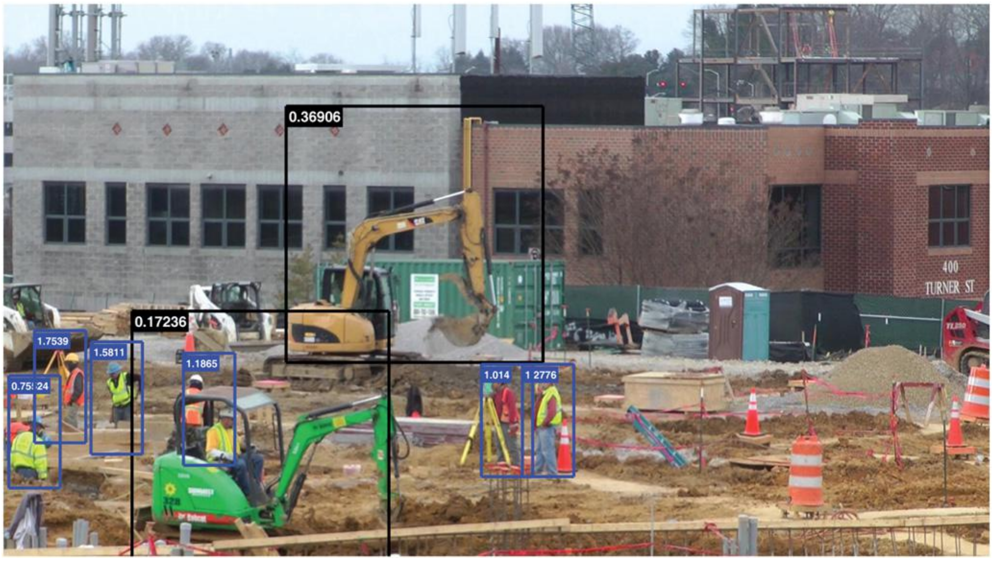

Fig. 24. Detection of excavators and construction workers in proximity of each other.

- Temporal reasoning for 2D detection of resources. Given the nature of construction, it is natural for resources to leave the field of view of a fixed camera and come back at a later time. Also there might be cases for which a resource is fully occluded temporally behind another static or dynamic resource on a jobsite. In both of these cases, there is a need for a temporal reasoning for the detection of the resources.
- Resource detection and localization using mobile cameras. The ability to detect construction workers and equipment from moving cameras opens exciting opportunities for context awareness. For example, a camera mounted on an excavator can minimize the chances of accidents by eliminating the blind spots and alert the operators about the detection of other resources in their proximities. Nonetheless moving cameras can create several dynamic changes in pose and configuration of other resources in 2D video streams. More research is needed on detecting resources using mobile cameras.

7. Conclusion

In this paper, we presented a novel method for automated 2D detection of construction workers and equipment from site video streams based on using histograms of oriented gradients and Hue–Saturation colors. Our results with average performance accuracies of 98.83%, 82.10%, and 84.88% for workers, excavators, and dump trucks respectively, hold the promise of applicability of the proposed method for first step of automated performance assessments. As validated, adding histogram of Hue–Saturation colors to oriented gradients significantly improved the detection of resources. Nonetheless, the detection of equipment and workers does not need these resources to have distinct colors as the HOG component of the histograms can easily represent generic cases. We also evaluated the effect of different model parameters (e.g., detector window size, cell size, and number of bins in histogram of colors) on detection accuracy. The proposed multi-scale sliding detection window is independent to scale and viewpoint of resources, as well as illumination conditions, and can detect resources while they are idle. Despite the good performance of HOG + C descriptors, they suffer from one major problem: high computation time. Sliding detection windows are relatively slow and hence unattractive for real-time applications necessary for many safety analysis purposes. Our future work involves implementing the HOG + C based sliding detection window algorithm using the NVIDIA CUDA parallel computing framework which

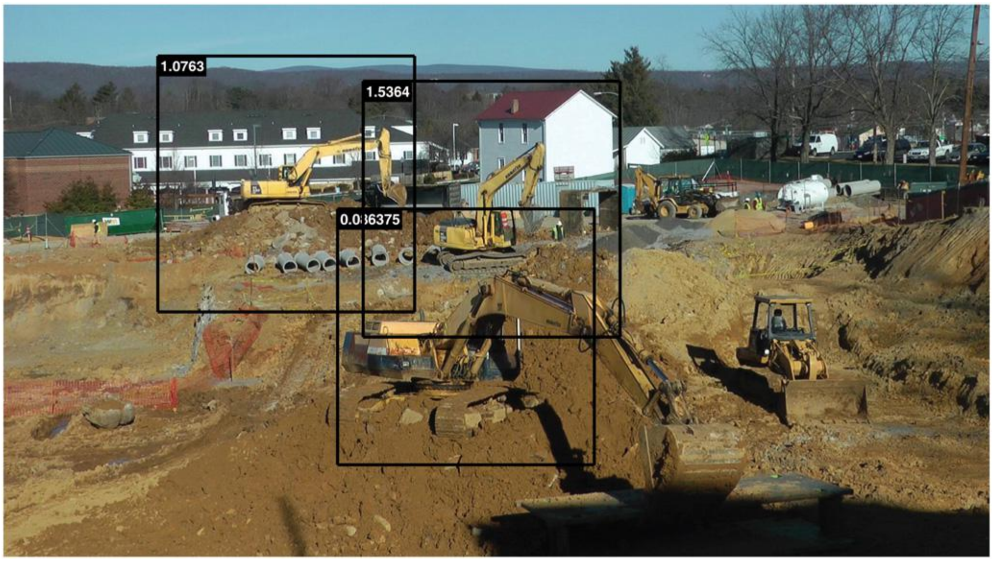

Fig. 25. Example of the capability of our proposed method in detection of multiple excavators with different viewpoints and distances to the camera.

can help achieve a real-time performance. Moreover, future work includes more exhaustive training and testing and also including different types of equipment, as well as varying body posture for the workers. Algorithmic development for the detection of resources across multiple video cameras, in addition to creating a temporal reasoning for those resources that leave a camera's field of view, or are fully occluded is currently under study.

Acknowledgment

The authors thank Arsalan Heydarian, Ao Chen, Moshe Zelkowicz, Fabian Capra, Rafael Suriel, and other members of the RAAMAC lab for their help with data collection. The support of Skanska, Holder Construction and Virginia Tech Facilities with data collection and the financial support of Virginia Tech's Institute of Critical Technologies and Applied Science (ICTAS) and Occupational Safety and Health Research Center (OSHRC) are also appreciated. This work is also partly supported by “el Patrimonio Autonomo Fondo Nacional de Financiamiento para la Ciencia, la Tecnologia y la Innovacion, Francisco Jose De Caldas” under contract RC No. 0394-2012 with Universidad del Norte.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.autcon.2012.12.002.

References

[1] E. Rezazadeh Azar, B. McCabe, Automated Visual Recognition of Dump Trucks in Construction Videos, J. Comput. Civ. Eng. 26 (6) (2012) 769–781, http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000179.

[2] H. Bay, A. Ess, T. Tuytelaars, L.v. Gool, SURF: speeded up robust features, Computer Vision and Image Understanding 110 (3) (2008) 346–359.

[3] I. Brilakis, M. Park, G. Jog, Automated vision tracking of project related entities, Advanced Engineering Informatics 25 (2011) 713–724.

[4] C.H. Caldas, D. Grau, C.T. Haas, Using global positioning system to improve materials-locating processes on industrial projects, ASCE Journal of Construction Engineering and Management 132 (7) (2006) 741–750.

[5] T. Cheng, M. Venugopal, J. Teizer, P.A. Vela, Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments, Automation in Construction 20 (2011) 1173–1184.

[6] S. Chi, C.H. Caldas, Automated object identification using optical video cameras on construction sites, Computer-Aided Civil & Infrastructure Engineering 26 (2011) 368–380.

[7] J.C. Christopher, A. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998) 121–167.

[8] N. Dalal, Finding people in images and videos, PhD Dissertation Institute National Polytechnique De Grenoble (2006).

[9] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. IEEE CVPR, San Diego, CA, 2, 2005, pp. 886–893.

[10] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: Proc. ECCV, 2, 2006, pp. 428–441.

[11] S. El-Omari, O. Moselhi, Integrating automated data acquisition technologies for progress reporting of construction projects, in: 26th International Symposium on Automation and Robotics in Construction, Austin TX, 2009.

[12] ENR, Don't blame the workers, engineering news-record, 2011.

[13] E. Ergen, B. Akinci, R. Sacks, Tracking and locating components in precast storage yard utilizing RIFD technology and GPS, Automation in Construction 16 (3) (2007) 354–367.

[14] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, TPAMI 32 (9) (2010) 1627–1645.

[15] R.J. Fontana, S.J. Gunderson, Ultra-wideband precision asset location system, in: IEEE Conference on Ultra Wideband Systems and Technologies, Baltimore MD, 2002.

[16] M. Golparvar-Fard, F. Pena-Mora, C.A. Arboleda, S. Lee, Visualization of construction progress monitoring with 4D simulation model overlaid on time-lapsed photographs, ASCE Journal of Computing in Civil Engineering 23 (6) (2009) 391–404.

[17] M. Golparvar-Fard, F. Pena-Mora, S. Savarese, Application of D4AR — a 4-dimensional augmented reality model for automating construction progress monitoring data collection, processing and communication, ITcon 14 (2009) 129–153.

[18] M. Golparvar-Fard, F. Pena-Mora, S. Savarese, Integrated sequential as-built and as-planned representation with D4AR tools in support of decision-making tasks in the AEC/FM industry, ASCE Journal of Construction Engineering and Management 137 (12) (2011) 1099–1116.

[19] J. Gong, C.H. Caldas, Computer vision-based video interpretation model for automated productivity analysis of construction operations, ASCE Journal of Computing in Civil Engineering 24 (2010) 252–263.

[20] J. Gong, C.H. Caldas, Construction site vision workbench: a software framework for real-time process analysis of cyclic construction operations, in: Proc. ASCE Int. Workshop on Computing in Civil Engineering, Auton, TX, 2009, pp. 64–73.

[21] J. Gong, C.H. Caldas, Data processing for real-time construction site spatial modeling, Automation in Construction 17 (2008) 526–535.

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

[23] P. Goodrum, C.T. Haas, C.H. Caldas, D. Zhai, J. Yeiser, D. Homm, Model to predict the impact of a technology on construction productivity, ASCE Journal of Construction Engineering and Management 137 (678) (2011).

[24] D. Grau, C.H. Caldas, C.T. Haas, P.M. Goodrum, J. Gong, Assessing the impact of materials tracking technologies on construction craft productivity, Automation in Construction 18 (2009) 903–911.

[25] J. Teizer, D. Lao, M. Sofer, Rapid automated monitoring of construction site activities using ultra-wideband, in: Proc. of 24th Int. ISARC, Kerala, India, 2007, pp. 23–28.

[26] V.R. Kamat, M. Akula, Integration of global positioning system and inertial navigation for ubiquitous context-aware engineering applications, in: Proc. National Science Foundation Grantee Conference, Atlanta GA, 2011.

[27] I. Laptev, Improvements of Object Detection using Boosted Histograms, BMVC, 2006.

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

[29] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET Curve in Assessment of, Detection Task Performance, NIST, 1997.

[30] M. Memarzadeh, A. Heydarian, M. Golparvar-Fard, J. Niebles, Real-Time and Automated Recognition and 2D Tracking of Construction Workers and Equipment from Site Video Streams, Computing in Civil Engineering (June 2012) 429–436.

[31] R. Navon, Automated project performance control (appc) of construction projects, Automation in Construction 14 (4) (2005) 467–476.

[32] R. Navon, R. Sacks, Assessing research in automated project performance control (APPC), Automation in Construction 16 (4) (2007) 474–484.

[33] NIST, Criteria for Performance Excellence, National Institute of Science and Technology, 2011–2012.

[34] C.H. Oglesby, H.W. Parker, G.A. Howell, Productivity Improvement in Construction, McGraw-Hill, New York, NY, 1989.

[35] M-W. Park, C. Koch, I. Brilakis, 3D Tracking of Construction Resources Using an On-site Camera System, Journal of Computing in Civil Engineering 26 (4) (2012) 541–549.

[36] M-W. Park, I. Brilakis, Construction worker detection in video frames for initializing vision trackers, Automation in Construction (ISSN: 0926-5805) 28 (2012) 15–25, http://dx.doi.org/10.1016/j.autcon.2012.06.001, (http://www.sciencedirect.com/science/article/pii/S0926580512001136).

[37] J. Song, C.H. Caldas, E. Ergen, C.T. Haas, B. Akinci, Field trials of RFID technology for tracking pre-fabricated pipe spools, in: Proc. of the 21st International Symposium on Automation and Robotics in Construction, 2004.

[38] J. Song, C.T. Haas, C.H. Caldas, Tracking the location of materials on construction job sites, ASCE Journal of Construction Engineering and Management 132 (9) (2006) 911–918.

[39] Y. Su, L. Liu, Real-time construction operation tracking from resource positions, in: ASCE Int. Workshop on Computing in Civil Engineering, Pittsburg, PA, 2007, pp. 200–207.

[40] P. Viola, M. Jones, Rapid object detection using boosted cascade of simple features, in: IEEE CVPR, Kauai HI, 1, 2001, pp. 1–9.

[41] J.v.d. Weijer, C. Schmid, Coloring local feature extraction, in: Proc. ECCV, 2, 2006, pp. 332–348.

[42] J. Yang, P.A. Vela, J. Teizer, Z.K. Shi, Vision-based crane tracking for understanding construction activity, in: Proc. ASCE Int. Workshop on Computing in Civil Engineering, Miami FL, 2011, pp. 258–265.

[43] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: IEEE CVPR, 2011, pp. 1385–1392.

[44] J. Zou, H. Kim, Using hue, saturation, and value color space for hydraulic excavator idle time analysis, ASCE Journal of Computing in Civil Engineering 21 (2007) 238–246.

[45] D. Forsyth, J. Ponce, Computer Vision: A Modern Approach, Second edition Pearson Education Inc., 2011.