Automated 2D Detection and Localization of Construction Resources in Support of Automated Performance Assessment of Construction Operations

Milad Memarzadeh

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Civil and Environmental Engineering

Mani Golparvar-Fard (Chair)
Jesus M. de la Garza
Linsey C. Marr
Juan Carlos Niebles

12/10/12
Blacksburg, VA

Keywords: Histogram of Oriented Gradients, Support Vector Machine, HSV Colors, Resource Detection and Localization, Performance Monitoring, Deformable Part-based Models

Automated 2D Detection and Localization of Construction Resources in Support of Automated Performance Assessment of Construction Operations

Milad Memarzadeh

ABSTRACT

This study presents two computer vision based algorithms 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, two new algorithms based on Histograms of Oriented Gradients and Colors (HOG+C), 1) HOG+C sliding detection window technique, and 2) HOG+C deformable part-based model are proposed and their performance are compared to the state-of-the-art algorithm in computer vision community. Furthermore, a new comprehensive benchmark dataset containing over 8,000 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. The preliminary results with average performance accuracies of 100%, 92.02%, and 89.69% for workers, excavators, and dump trucks respectively, indicate the applicability of the proposed methods for automated activity analysis of workers and equipment from single video cameras. Unlike other state-of-the-art algorithms in automated resource tracking, these methods particularly detects idle resources and does not need manual or semi-automated initialization of the resource locations in 2D video frames.

Acknowledgements

I would like to thank my advisor, Professor Mani Golparvar-Fard, for his guidance, encouragement, and support over the course of my graduate career. I am highly grateful and proud to have worked under his supervision. He taught me a great deal besides sciences and engineering; mentorship, work ethics and patience. I would also like to thank my committee members for their participation and support. I would like to express my special thanks to Professor Juan Carlos Niebles for his helpful discussions and guidance. He has been an insightful and helpful committee member and teacher during my graduate studies.

I would like to thank all my group members in real-time and automated monitoring and control (RAAMAC) lab, previous and present, for their help, advice, encouragement, discussions, time, and support including Vahid Balali, Youngjib Ham, Ao Chen, Moshe Zelkowicz, Fabian Capra, Rafael Suriel, Ashish Sharma, and Chris Lazenby. Special thank goes to Arsalan Heydarian for leading and conducting primary data collection. The photos and video sequences are all taken by the author, Arsalan Heydarian, Moshe Zelkowicz, Fabian Capra, Rafael Suriel, Ashish Sharma, and Chris Lazenby in the span of summer 2011 to November 2012 and from the construction jobsites on the campus of Virginia Tech or around town of Blacksburg. The support of Skanska, Holder Construction and Virginia Tech Facilities with data collection and the financial support from Virginia Tech's Occupational Safety and Health Research Center (OSHRC), and Institute for Critical Technology and Applied Science (ICTAS) are also appreciated. Any opinions, findings, and conclusions or recommendations expressed in this material are those of author and do not necessarily reflect the views of the OSHRC and ICTAS.

Finally, I am forever grateful to my family: Meisam, and Mohsen for their continued love and support in my life endeavors. I am deeply indebted to my father and mother for their endless and unconditional support and love. At the end, I cannot ever forget my parents' sacrifices to provide me encouragement and motivation to pursue my goals.

Table of Contents

List of Figures

List of Tables

Preface/ Attribution

The thesis author was responsible for substantial contributions to the content and writing of the two co-authored manuscripts presented in Chapter 2 and 3. He played a lead role in writing these manuscripts and the rest of the thesis including the literature review, collecting data, and developing the algorithms.

The co-authors participated in the development and drafting of ideas and were equal partners with the thesis author in the review and revision of the manuscripts.

Chapter 1
Introduction and Literature Review

1.1 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 activity analysis and enable timely identification of productivity, safety, and occupational health issues. Continuous and systematic activity analysis in particular allows companies to identify solutions to minimize low operational efficiencies. Once these solutions are implemented, they could be followed up with additional video-based analyses to validate whether those solutions addressed the performance issue, or if there is still a need for further improvements. In addition to their immediate benefits, 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 at their entirety 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 image 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, the first key challenge 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, many challenging problems remain open.

As a step towards fully automated performance assessment methods, this study focuses on the problem of automated 2D detection of workers and equipment in site video streams and a number of applications this enables. Figure 1.1 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.

Figure 1.1. 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.

The new methods proposed in this research project, expand on the work originally presented in Dalal and Triggs (2005) and Felzenszwalb et al. (2010) with addition of several novel components to the algorithm that significantly improve the performance of the methods. It will also accompany with exhaustive 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 will also be provided. The perceived benefits and limitations of the proposed method in the form of open research challenges are presented. More details about this project can be find at www.raamac.cee.vt.edu/detectiontracking.

1.2 Background

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 (Oglesby et al. 1989), 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 (Gong and Caldas 2011, Su and Liu 2007) and can be subjective (Golparvar-Fard et al. 2011). The significant amount of information which needs to be collected may also adversely affect the quality of the analysis (Golparvard-Fard et al. 2009, Gong and Caldas 2009). Such limitations minimize the opportunities for continuous benchmarking and monitoring which is a key element in performance improvement (NIST 2011-2012). Hence, many critical decisions may be made based on incomplete or inaccurate information, ultimately leading to project delays and cost overruns. 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.

1.2.1 3D Localization and Tracking of Construction Resources Using Sensors

In recent years, a number of research studies (Gong and Caldas 2011, Su and Liu 2007, Gong and Caldas 2010, Goodrum et al. 2011) 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 (El-Omari and Moselhi 2009, Ergen et al. 2007, Grau et al. 2009, Navon and Sacks 2007, Song et al. 2004, Song et al. 2006), Ultra WideBand (UWB) (Cheng et al. 2011, Teizer et al. 2007), 3D range imaging

cameras (Gong and Caldas 2008), global and local positioning systems (GPS) (Teizer et al. 2007, Gong and Caldas 2008), and computer vision techniques (Brilakis et al. 2011, Park et al. 2011) 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 localization even in the presence of severe multipath (Fontana and Gunderson 2002). In a recent case, Teizer et al. (2007) 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 (Brilakis et al. 2011). As such, the implementation of this system may be challenged and can be costly where there 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 (Gong and Caldas 2008) and resource tracking (Teizer et al. 2007) 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 (Caldas et al. 2006). 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 (Kamat and Akula 2011). 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 (Brilakis et al. 2011). As a result, several research studies (El-Omari and Moselhi 2009, Ergen et al. 2007, Navon and Sacks 2007, Navon 2005) 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.

1.2.2 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 (Oglesby et al. 1989). 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 (ENR 2011) or video-based activity analysis (Oglesby et al. 1989), 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 (Gong and Caldas 2011, Golparvar-Fard et al. 2009, Navon and Sacks 2007, Brilakis et al. 2011, Golparvar-Fard et al. 2009) 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 and tracked in 2D. Recently developed methods (Gong and Caldas 2011, Zou and Kim 2007) are either simulated in controlled environments or have primarily focused on automating the 3D tracking assuming semi-automated detection of resources in 2D. Others such as (Brilakis et al. 2011, Park et al. 2011, Yang et al. 2011) use *a priori* knowledge for their assessments such as expected known locations for tracking a tower crane (Yang et al. 2011), or application of Scale Invariant Feature Transforms (SIFT) (Lowe 2004) and Speeded Up Robust Features (SURF) (Bay et al. 2008) for initial recognition.

Two recent works (Azar and McCabe 2011, Chi and Caldas 2011) focus on developing techniques for automated 2D detection and localization of construction workers and equipment. Particularly, (Chi and Caldas 2011) 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 (Azar and McCabe 2011). Particularly, the application of HOG detectors (Dalal

and Triggs 2005), Haar-like detectors (Viola and Jones 2001), Haar-HOG cascade (Bay et al. 2008), and Blob-HOG cascade methods are proposed. Due to the application of background subtraction, these methodologies are not able to recognize idle resources.

In the computer vision community, there is a large number of emerging works in the area of human detection and pose estimation (Dalal and Triggs 2005, Felzenszwalb et al. 2010, Dalal et al. 2006, Laptev 2006, Yang and Ramanan 2011). The results of these algorithms seem to be both effective and accurate. For example, (Felzenszwalb et al. 2010) 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 2006) 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 exhaustively 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.

1.2.3 Object Detection and Pose Estimation Using Part-based Models

Part-based models have appeared in recent years under various formalisms. The basic premise is that objects can be modeled as a collection of local template that deform and articulate with respect to one another.

In AEC community, the first research project that has focused on application of part-based models is Azar and McCabe (2012) that helps improve the detection of construction resources. They have proposed the object recognition system based on mixtures of appearances of deformable body parts of the hydraulic excavator. Their proposed approach shows promising results, however has few major drawbacks. Firstly, they have used the arm of excavator as a root which limits the applicability of the proposed approach for detection of whole excavator which is more important than detecting parts (i.e. the detection only happen on the arm). Moreover, they only used one part which may not enough for detecting construction resources in dynamic and highly occluded construction environments. Their proposed algorithm is limited to detection of excavators only from side views and cannot detect excavator in front and rear views which is not applicable for continuous monitoring of earthmoving operations. Moreover, their results shows that the accuracy of the part-based model in 2D detection of excavators are less than the state-of-the-art HOG approach, while in the proposed methods in this research project the accuracy of detection using part-based models outperforms the state-of-the-art HOG.

In the computer vision community there are a lot of researches in recent years on developing the part-based models for object detection and pose estimation. Felzenszwalb et al.

(2008, 2010) proposed an object detection system based on mixtures of multi-scale deformable part models. Their weakly supervised training procedure requires only the bounding box around the whole object. Their algorithm shows the promising results in object detection in 2D static images. Sun and Savarese (2011) proposed an articulated part-based model for joint object detection and pose estimation. This method represents an object as a collection of parts at multiple level of detail, from coarse-to-fine, where parts at every level are connected to a coarser level through a parent-child relationship. Pepik et al. (2012) have recently extended the successful discriminatively trained deformable part models (Felzenszwalb et al. 2008, Felzenszwalb et al. 2010), to include both estimate of viewpoint and 3D parts that are consistent across viewpoints. In fact, they added the 3D geometry to deformable part models which can be used for 3D scene understanding or 3D object tracking. Yang and Ramanan (2011) have proposed a method for human pose estimation in static images based on part-based models. They presented a general, flexible mixture model for capturing contextual co-occurrence relations between parts, augmenting standard spring models that encode spatial relationships.

All of the abovementioned researches in the computer vision community are tested on the controlled environment, which is not the case in dynamic and occluded construction jobsites. Moreover, the proposed approach in this research project expands on the part-based models by adding the histogram of colors which significantly improves the detection accuracy in saturated environments like construction jobsites.

1.3. Research Objectives

This research focuses on algorithmic development of computer vision and machine learning techniques for automated detection of various construction resources in support of automated

performance assessment of construction operations. Particularly, the objective is to create, develop, and validate two new algorithms for automated 2D detection and localization of construction resources from site video streams and compare their performance to state-of-the-art in computer vision community. These propose methods would,

- Detect and localize multiple resources in 2D video streams automatically (could be real-time), and
- Generate the deformable part-based model for detection of resources under high degrees of occlusion (suitable for activity analysis)

The outcome of proposed approach could provide useful visual information about the various resources that are operating/working on the jobsite for further analysis such as 3D tracking, action recognition, and performance assessment.

Chapter 2
Proposed Methodology

Some of the materials in Chapter 2 are from two submitted co-authored manuscripts (Memarzadeh et al. 2012a,b) listed in references.

2. Vision-based Detection of Resources

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 overall flowchart of the proposed approaches is illustrated in Figure 2.1.

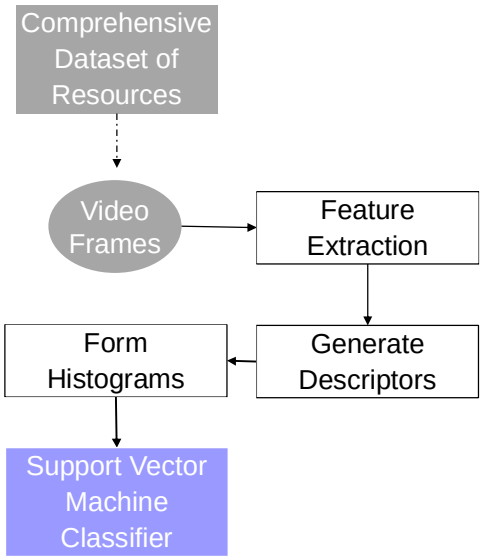

Figure 2.1. The flowchart of the proposed methods

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.

2.1. HOG+C Sliding Detection Window Technique

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 (Dalal and Triggs 2005) with Histograms of Color (HOC) to create an automated 2D detection method. These steps are described in the following subsections.

2.1.1 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. 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.

2.1.2 Resource Detection and Classification for each Detection Window

Figure 2.2 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 Figure 2.2); and 2) color cues captured by a HOC descriptor (right side of Figure 2.2). 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.

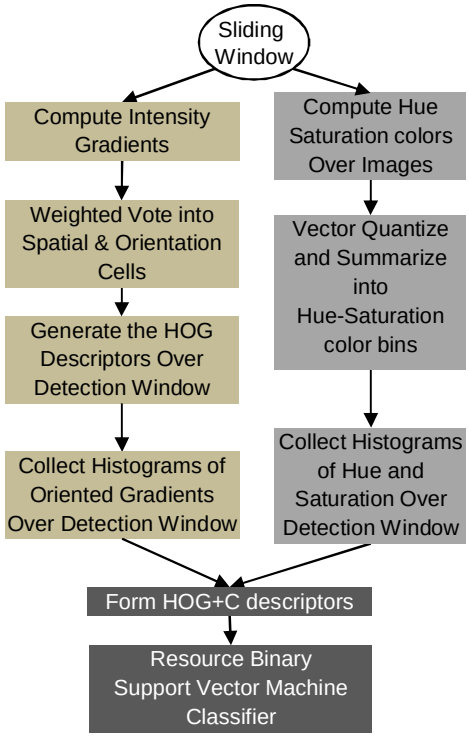

Figure 2.2. The algorithm for automated detection of construction resources

2.1.3 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 (Dalal and Triggs 2005). 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, the detection window (Figure 2.3a) is divided into $t_x \times t_y$ local spatial regions (*cells*) where each cell contains $u \times v$ pixels (Figure 2.3b). 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: $i \times \pi / n \;\; i = 1,...,n$ (see Figure 2.3c). 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 (Dalal 2006), 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 (Felzenszwalb et al. 2010), 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 (2005), Felzenszwalb et al. (2010) showed that the performance of the HOG descriptors could be improved; which is the rationale behind their application in our method.

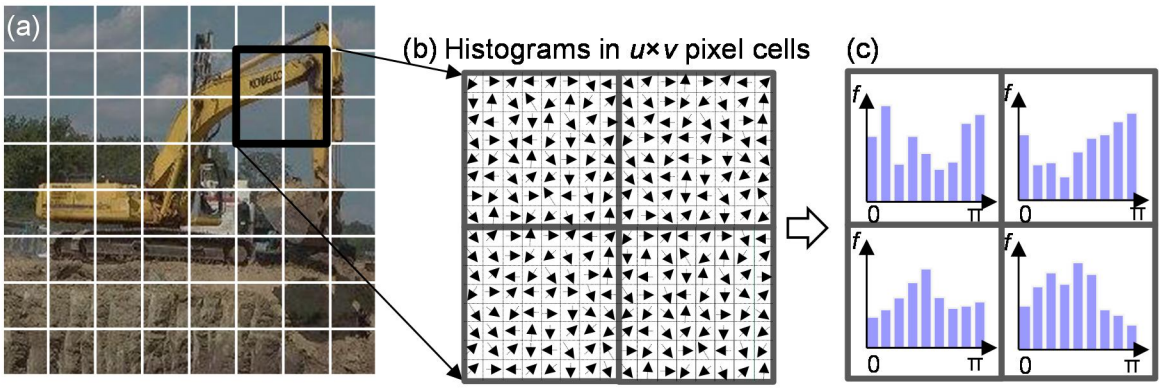

Figure 2.3. The formation of the Histogram of Oriented Gradients for each detection window: (a) a 250×250pixel detection window, (b) 4×4 pixel cells in each window, and (c) the Histogram of Oriented Gradients corresponding to 4 cells.

2.1.4 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 (2006), instead of using Red-Green-Blue (RGB) color space, in our algorithm, we use Hue-Saturation-Value (HSV) colors (Forsyth and Ponce 2011). 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. 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.

2.1.5 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 (Christopher and Burges 1998). Given *n* labeled training datapoints $\{x_i, y_i\}$, wherein x_i ($i = 1,...,n, x_i \in R^d$) 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 hyper-plane $\mathbf{w}^T\mathbf{x} + b = 0$ between the positive and negative samples. We assume that there is no *a prior* 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:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{2}$$

$$subject\ to: y_i(w.x_i + b) \geq 1 \quad for \quad i = 1,...,N$$

The presence of noise and occlusions which is typical in construction site video streams produces outliers in the SVM classifiers. Hence the slack variables ξ_i are introduced and consequently the SVM optimization problem can be written as:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \tag{3}$$

$$subject\ to: y_i(w.x_i + b) \geq 1 - \xi_i \quad for\ \ i = 1,...,N$$

$$\xi_i \geq 0 \ \ for\ i = 1,...,N$$

In this formula, *C* represents a penalty constant that can be determined by a cross-validation technique. As observed in Figure 2.4, 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.

Data: Training sets $(x_1, l_1),...,(x_n, l_n)$ 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	
1	**for** each positive and negative example file list
2	**for** each cell in an example
3	create HOG descriptor
4	create Hue-Saturation color descriptor
5	concatenate into HOG+C descriptor
6	**end for**
7	append to the train_positive and train_negative lists
8	**end for**
9	**find** support vectors
10	**return** *M*

Figure 2.4. Algorithm for training process

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. Figure 2.5 shows the algorithm for the detection (testing) process of a resource classifier.

Data: Learned one-against-all SVM Model (M_j) per resource
Testing images and video frames
Results: bounding boxes and classification scores per box

```
1   for each image in the testing list
2     for each spatial scale in search scale
3       resample image to new scale
4       for each detection window in an image
5         compute HOG descriptor
6         compute Hue-Saturation color descriptor
7         concatenate and from the HOG+C descriptor
8         analyze the detection window with Mj
10      end for
12    end for
13    non-maximum suppression
14  end for
15  return bounding boxes and classification scores
```

Figure 2.5. Algorithm for detection (testing) process

2.2 HOG+C Deformable Part-based Model

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 weakly supervised in the sense that we annotate bounding boxes around each resource in the positive images, however the locations of parts are not annotated and will be automatically learned. (This means the same dataset used in the previous section could be used for training/testing in this step). Each example x is scored by a function of the form, $f_\beta(x) = \max_z \beta.\Phi(x,z)$, where β is a vector of model parameters and z are latent values (e.g. the part placements). Latent Support Vector Machine (LSVM) is used to learn the model per resource. During the testing stage, the proposed method automatically places the bounding boxes around the entire object and identifies different parts, 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, 2) HOG+C descriptors which are formed by concatenating HOG (Memarzadeh et al. 2012b) with Histograms Of Color (HOC) to create an automated 2D detection method (Memarzadeh et al. 2012a), 3) mixture of models which smooth the problem of intra-class variability, and 4) deformable part-based models to improve the performance and handle the occlusions.

2.2.1 Overview

Figure 2.6 shows the overview of the proposed methodology.

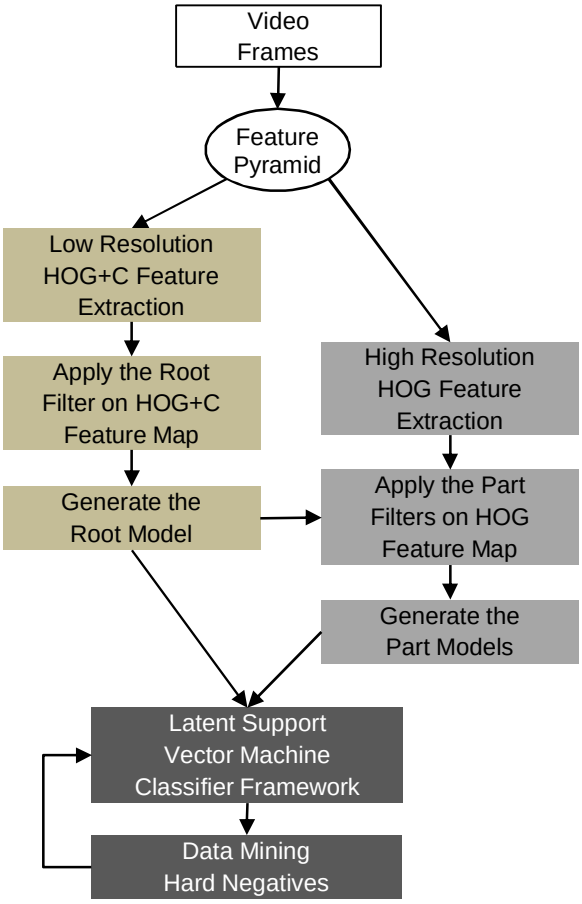

Figure 2.6. The overview of the proposed methodology

2.2.2 Features Representation

For representing the root filter, we use the HOG+C feature extraction process explained in Memarzadeh et al. (2012a) on the coarse gradients on the low-resolution top level of feature pyramid (see Figure 2.7). The part filters are placed several levels down in the feature pyramid so the resolution of the image would be twice as the resolution of the root filter. Using higher resolution features for defining part filters is essential for obtaining high recognition performance (Felzenszwalb et al. 2010).

2.2.3 Models

The models involve linear filters which are rectangular templates specifying weights for subwindows of a feature pyramid. Let F be a $w \times h$ filter and H be a feature pyramid and $p = (x, y, l)$ specify a position (x, y) in the l-th level of the pyramid, and $\phi(H, p, w, h)$ denote the vector obtained by concatenating the feature vectors in the $w \times h$ subwindow of H with top-left corner at p in row-major order. The score of F at p is $F'. \phi(H, p, w, h)$, where F' is the vector obtained by concatenating the weight vectors in F in row-major order.

2.2.4 Deformable Part Models

The proposed models are defined by a coarse HOG+C root filter that approximately covers an entire object and higher resolution part HOG filters that cover smaller parts of the object. Figure 2.7 illustrates the placement of root and part models in a feature pyramid. The root filter location defines a detection window. The part filters are placed several levels down in the pyramid, so the features at that level are computed at twice the resolution of the feature in the root filter level. Consider building a model for an excavator. The root filter cold capture coarse resolution edges such as the equipment body boundary while the part filters could capture details such as bucket, arm, and truck.

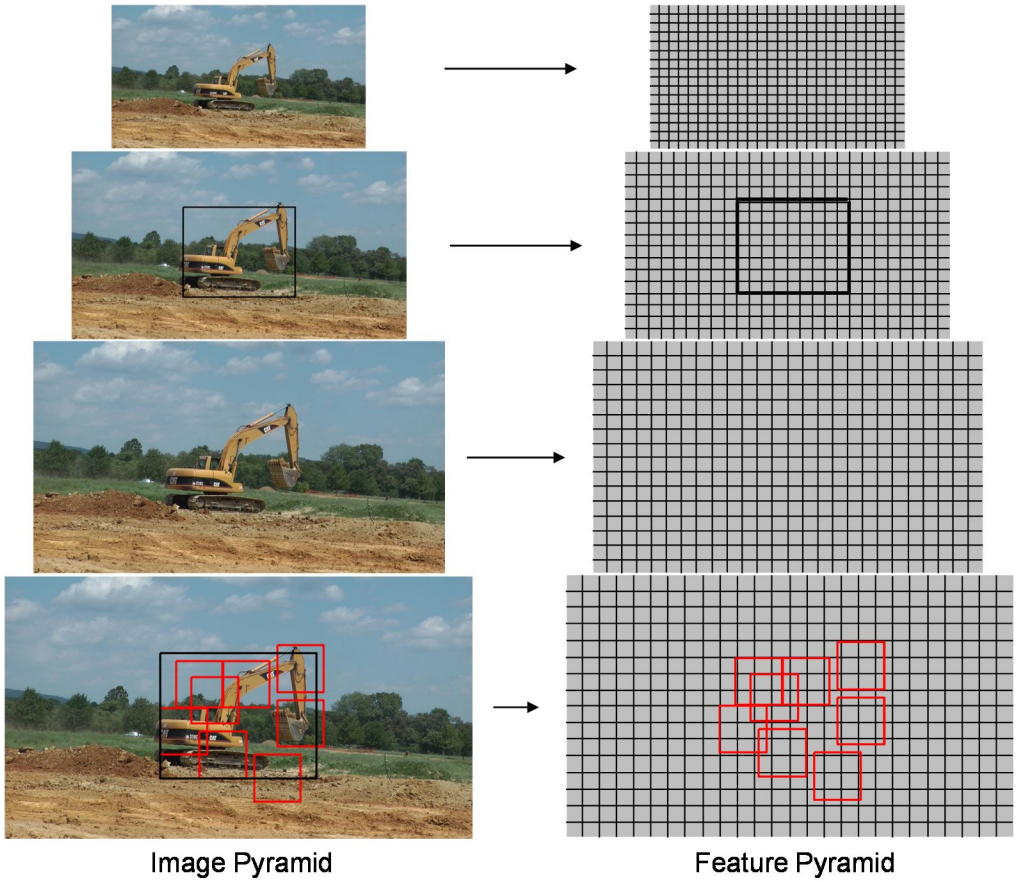

Figure 2.7. Illustration of Image and Feature Pyramid; The part filters are placed several levels down in the pyramid, so that the features at that level are computed at twice resolution of the feature in the root filter level

A model for a resource with n parts is formally defined by a $(n+2)$-tuple $(F_0, P_1, ..., P_n, b)$, where F_0 is a root filter, P_i is a model for the i-th part, and b is a real-valued bias term. Each part model is defied by a 3-tuple (F_i, v_i, d_i), where F_i is a filter for the i-th part, v_i is a two-dimensional vector specifying an "anchor" position for part i relative to the root position, and d_i is a four-dimensional vector specifying coefficients of a quadratic function defining a deformation cost for each possible placement of the part relative to the anchor position. Deformation costs are like springs connecting each part filter to the root filter, leading to a star-structured model.

A resource hypothesis specifies the location of each filter in the model in a feature pyramid, $z = (p_0, ..., p_n)$, where $p_i = (x_i, y_i, l_i)$ specifies the level and position of the i-th filter. The score of a hypothesis is given by the scores of each filter at their respective locations minus a deformation cost that depends on the relative position of each part with respect to the root (the spatial prior), plus the bias (Felzenszwalb et al. 2010),

$$score(p_0, ..., p_n) = \sum_{i=0}^{n} F_i' . \phi(H, p_i) - \sum_{i=1}^{n} d_i . \phi_d(dx_i, dy_i) + b \tag{4}$$

where

$$(dx_i, dy_i) = (x_i, y_i) - (2(x_0, y_0) + v_i) \tag{5}$$

gives the displacement of the i-th part relative to its anchor position and

$$\phi_d(dx, dy) = (dx, dy, dx^2, dy^2) \tag{6}$$

are deformation features.

The score of a hypothesis z can be expressed in terms of a dot product, $\beta . \psi(H, z)$, between a vector of model parameters β and a vector $\psi(H, z)$, which are as follow,

$$\beta = (F_0', ..., F_n', d_1, ..., d_n, b) \tag{7}$$

$$\begin{aligned} \psi(H, z) = (&\phi(H, p_0), ..., \phi(H, p_n), \\ &-\phi(dx_1, dy_1), \ ... \ , -\phi(dx_n, dy_n), 1) \end{aligned} \tag{8}$$

This illustrates a connection between the models and linear classifiers. This relationship is used for learning the model parameters with LSVM framework.

To detect objects in an image, we compute an overall score for root locations according to the best possible placement of the parts,

$$score(p_0) = \max_{p_1,\ldots,p_n} score(p_0,\ldots, p_n) \tag{9}$$

High-scoring root locations define detections while the locations of the parts that yield a high-scoring root location define a full resource hypothesis. This process is illustrated in Figure 2.8.

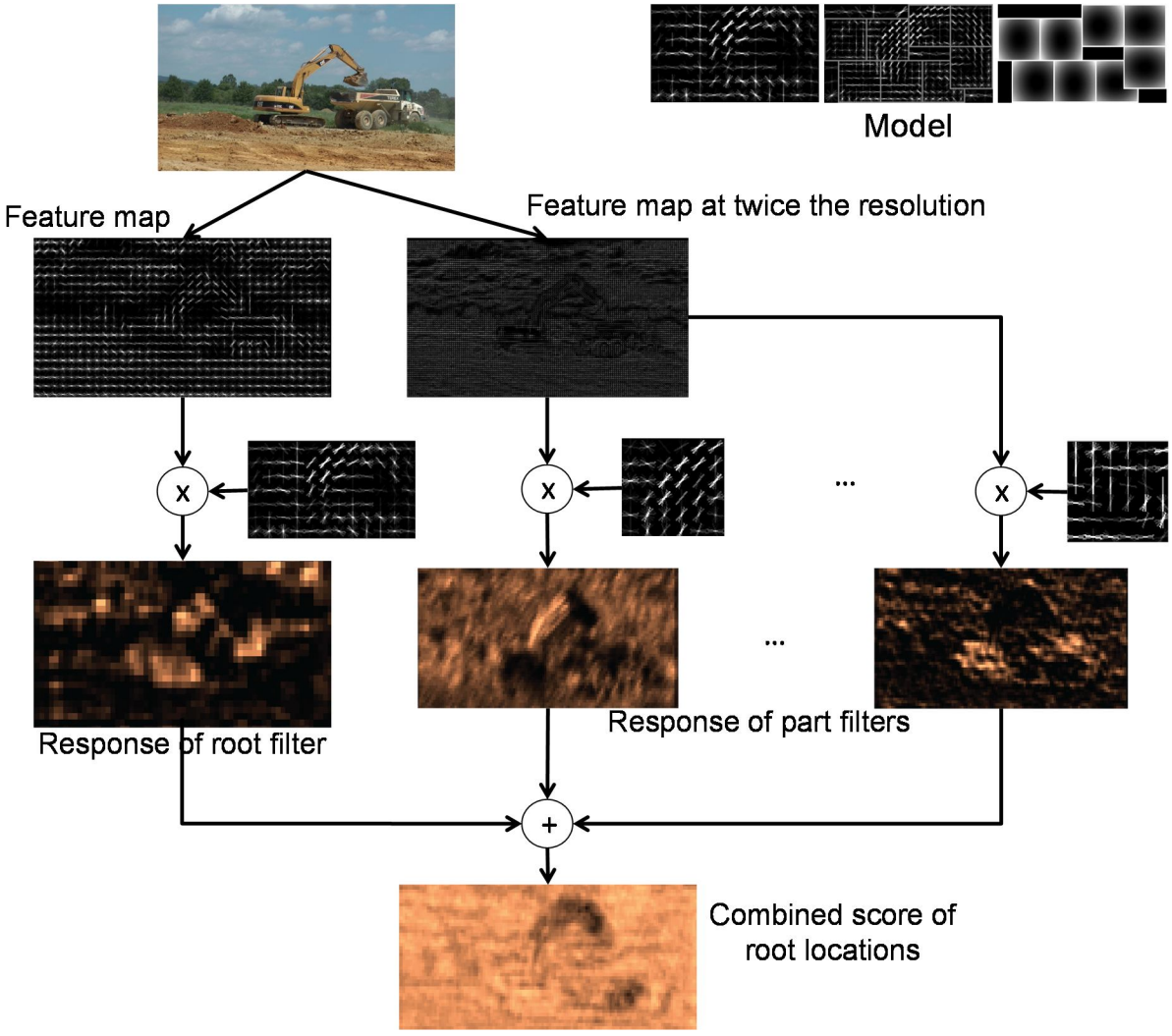

Figure 2.8. Responses of the root and part filters are computed a different resolution in the feature pyramid. The responses are combined to yield a final score for each root location. We show the responses for the "arm" and "truck" parts.

2.2.5 Learning

The RAAMAC-2012 dataset consists of a large set of video frames with bounding boxes around each instance of a resource. Let $D = (\langle x_1, y_1 \rangle, \ldots, \langle x_n, y_n \rangle)$ be a set of labeled examples where $y_i \in \{-1,1\}$ and x_i specifies a HOG+C feature pyramid, $H(x_i)$, together with a range, $Z(x_i)$, of valid placements for the root and part filters. For each positive example in the training set, $Z(x_i)$ is defined so the root filter must be placed to overlap the bounding box by at least 50%. Negative examples come from images that do not contain the specific resource form construction jobsite (e.g., construction site background, other resources).

2.2.5.1 Latent Support Vector Machine (LSVM)

The proposed classifier scores an example x with a function of the form (Felzenszwalb et al. 2010),

$$f_\beta(x) = \max_{z \in Z(x)} \beta . \Phi(x, z) \tag{10}$$

where β is a vector of model parameters and z are latent values. The set $Z(x)$ defines the possible latent values for an example x. A binary label for x can be obtained by thresholding its score. Training β from labeled examples $D = (\langle x_1, y_1 \rangle, \ldots, \langle x_n, y_n \rangle)$, where $y_i \in \{-1,1\}$, is similar to classical SVM [49] by minimizing the following objective function,

$$L_D(\beta) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i f_\beta(x_i)) \tag{11}$$

where $\max(0, 1 - y_i f_\beta(x_i))$ is the standard hinge loss (slack variables) and the constant C controls the relative weight of the regularization term. It should be noticed that if there is a single

possible latent value each example ($|Z(x_i)|=1$, i.e., only root filter) then f_β is linear in β, and the learning would be exactly the same as linear SVM.

2.2.6 Data Mining Hard Negatives

In the training a model, the number of negative examples is very large (a single image can yield 10^5 examples for a sliding detection window classifier). This can make the training process infeasible regard to computation time. Instead, it is common to construction training data consisting of the positive instances and "hard negative" instances, where the hard negatives are data-mined from the very large set of possible negative examples. The main idea is that the proposed method trains a model with an initial subset of negative examples, and then collects negative examples that are incorrectly classified by this initial model to form a set of hard negatives. A new model is trained with the hard negative examples and the process may be repeated a few times.

The hard negatives of D relative to β is defined as (Felzenszwalb et al. 2010),

$$M(\beta, D) = \{\langle x, y\rangle \in D \mid y f_\beta(x) \leq 1\} \tag{12}$$

where $M(\beta, D)$ are training examples that are incorrectly classified or near the margin of the classifier defined by β.

2.2.7 Non-Maximum Suppression

Detecting resources in video frames usually results to multiple overlapping detections of each instance of a resource. The greedy procedure is used for eliminating repeated detections via non-

maximum suppression. After applying the testing process, we have a set of detections D for a particular resource. Each detection is defined by a bounding box and a score. We sort the detections in D by score, and greedily select the highest scoring ones while skipping detections with bounding boxes that are at least 50% covered by a bounding box of a previously selected detection.

.

Chapter 3
Comprehensive Data Collection and Validation Metrics

Some of the materials in Chapter 2 are from submitted co-authored manuscript (Memarzadeh et al. 2012a) listed in references.

3.1 Data Collection

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. For this purpose, over 300 hours of video streams are collected from five ongoing construction projects: (1) Center for the Arts Construction Project; (2) Signature Engineering Building Construction Project; (3) Academic and Student Affairs Building Project; (4) Virginia Tech's Corporate Research Center Extension Project; and (5) the New River Resource Authority (NRRA) Area C Expansion Project (Heydarian 2011). These video streams are converted into images (around 10,000 images) with the bounding box information around the object of interest. 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 Figure 3.1, 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. This dataset is structured for both training and testing purposes so that it can be released to the community for further development and validation of new algorithms.

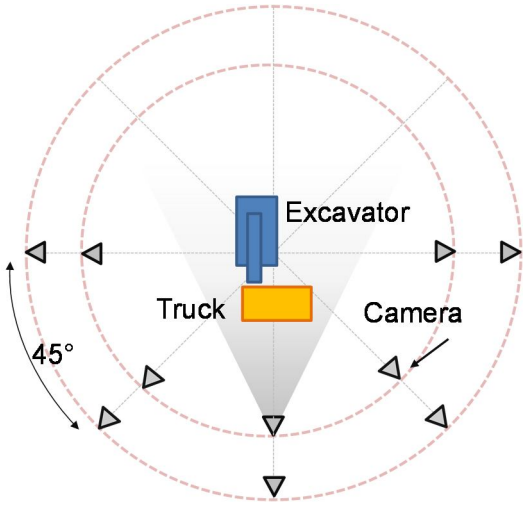

Figure 3.1. Data Collection and Experimental Setup

Table 3.1 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, (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.

Table 3.1. 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

3.2 Performance Evaluation Metrics

To quantify and benchmark the performance of the proposed 2D detection algorithms, 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) (Martin et al. 1997). Both of these 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:

3.2.1 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:

$$precision = \frac{TP}{TP + FP} \tag{13}$$

$$recall = \frac{TP}{TP + FN} \tag{14}$$

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 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.

3.2.2 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:

$$miss\ \ rate = 1 - recall\ \ rate = \frac{FN}{TP + FN} \tag{15}$$

$$FPPW = \frac{FP}{TN + FP} \tag{16}$$

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

Chapter 4
Experimental Results

4.1 Single Resource Detection

In the following sections, we first present the experimental results from our proposed algorithms. We also test the efficiency of our approaches on various model parameters. As observed in Figures 4.1 and 4.2, 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.

Figure 4.1. 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)

Figure 4.2. 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).

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 64bit platform with 24GB memory and 3.2GHz Core i7 CPU.

4.1.1 HOG+C Sliding Detection Window Technique

In our proposed HOG+C sliding detection window technique, the detectors have the following properties:

- The size 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 minutes.

Figure 4.3 shows a HOG+C descriptor which is learned using the worker training dataset. Figures 4.4 and 4.5, each show an example of a testing image, in addition to their HOG+C descriptors.

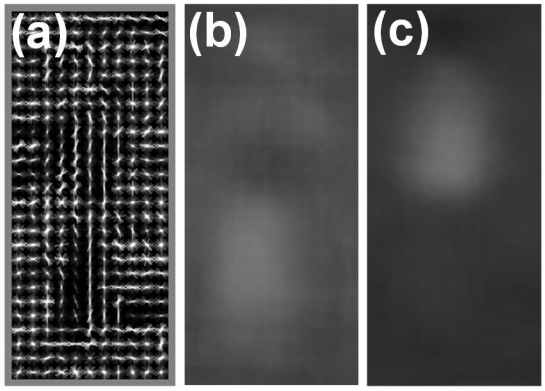

Figure 4.3. (a-c) The average oriented gradients, average hue values, and average saturation values over the worker dataset.

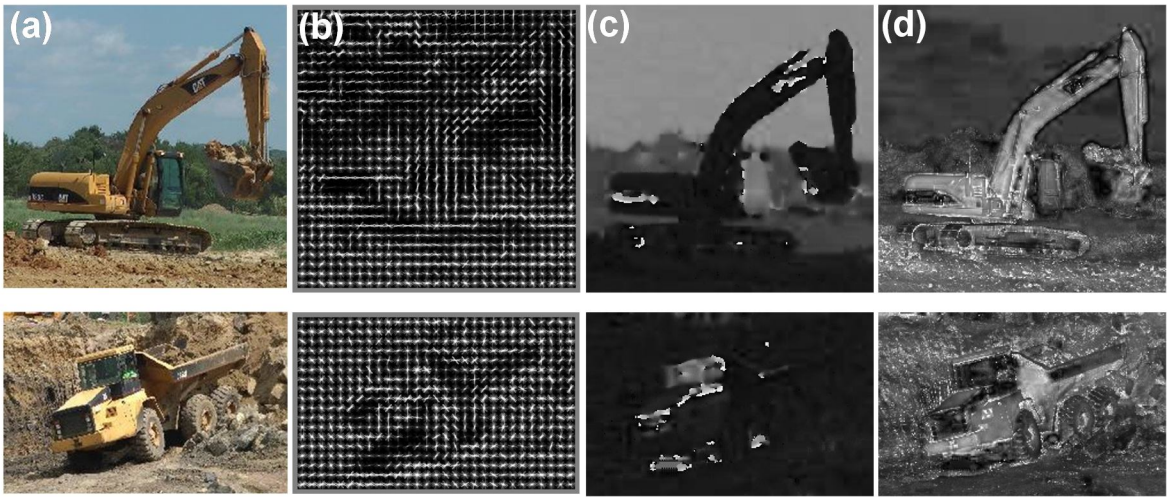

Figure 4.4. Example of testing for excavators and trucks datasets, respectively: each row: (a) a test image, (b) The oriented gradients, (c) Hue map, (d) Saturation map.

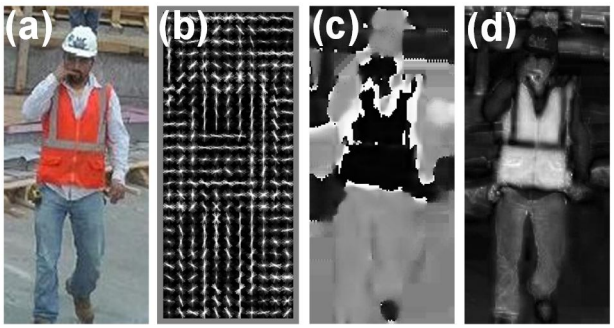

Figure 4.5. Example of testing for worker dataset: (a) a test image, (b) The oriented gradients, (c) Hue map, (d) Saturation map.

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. Figure 4.6 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. 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 4.1. 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.

Table 4.1. 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

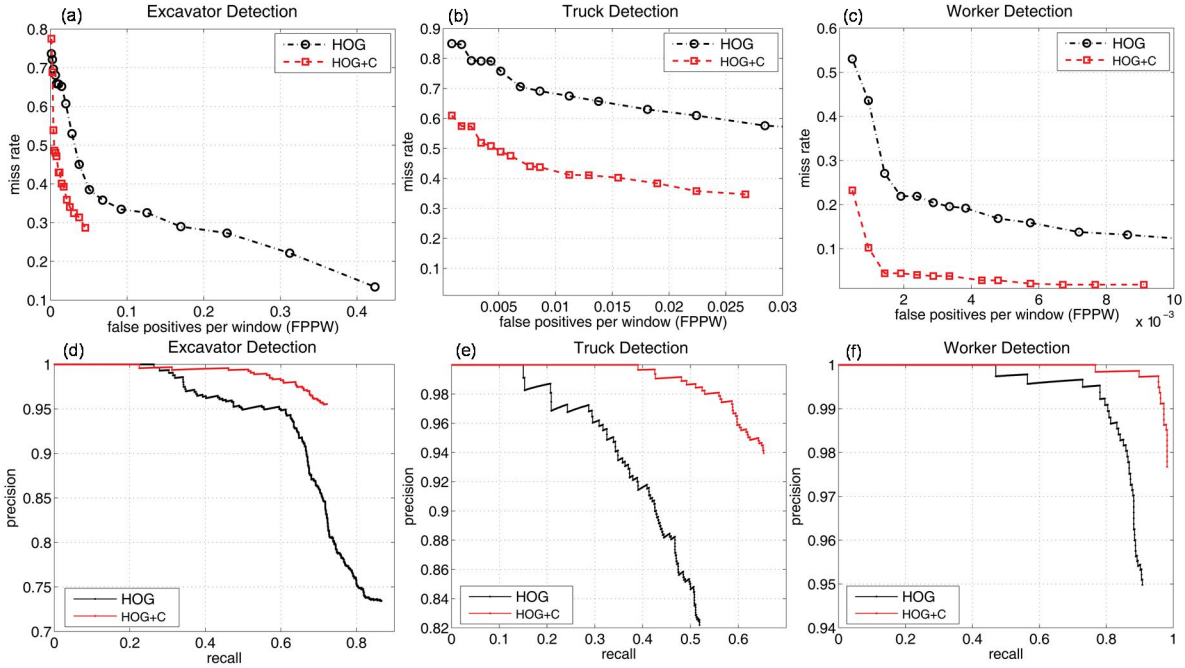

Figure 4.6. Overall results on performance of HOG and HOG+C on detection of construction resources. (a-c) DET and (d-f) precision-recall curves for detection of excavators, trucks, and workers, respectively.

Several examples of TP, FP, and FN for different resources detection methods are presented in Figure 4.7. As seen in the Figure 4.7a, 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. Figure 4.7b 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.

Figure 4.7. Example of TP, FP, and FN in detection of construction resources (the top left boxes show the classification scores).

a. 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 were selected based on the highest average performances and most reasonable computational times.

a.1 Effect of the Detection Sliding Window Size

Figures 4.8 and 4.9 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 Figures 4.8b, 4.9b), 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 Figures 4.8a, c, 4.9a, 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.

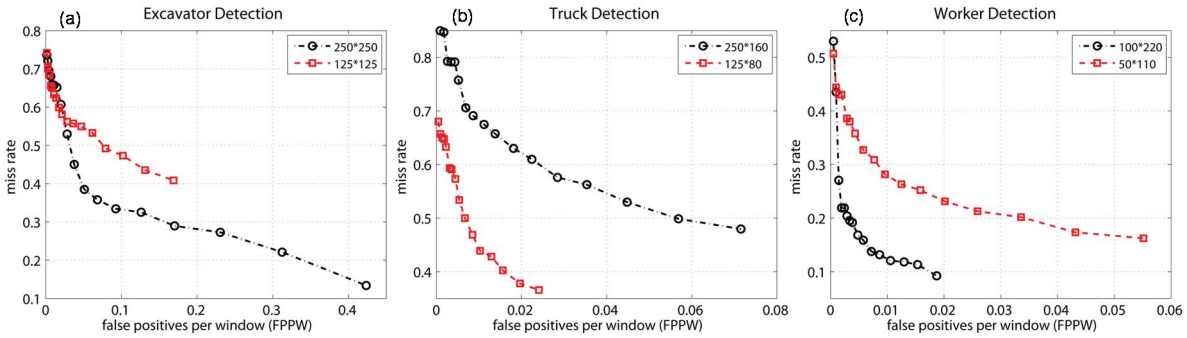

Figure 4.8. Effect of the detector window size on performance of HOG for detection of different construction resources.

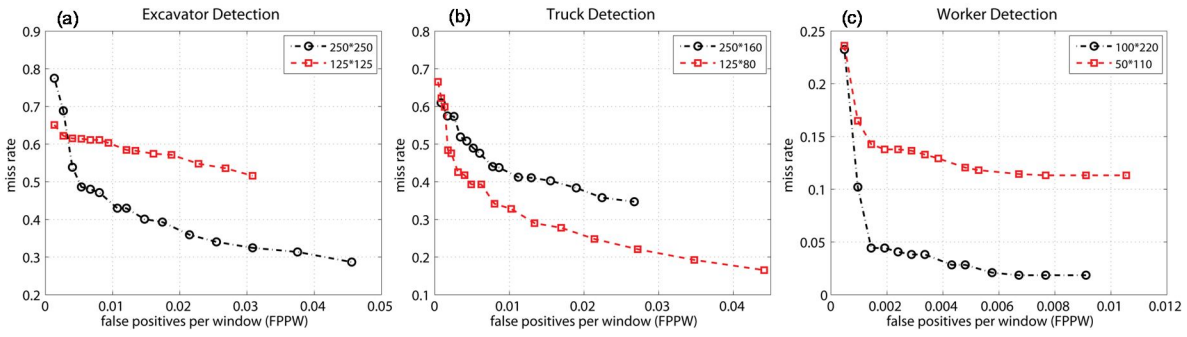

Figure 4.9. Effect of the detector window size on performance of HOG+C for detection of different construction resources.

a.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. Figures 4.10 and 4.11 demonstrate and compare the performance of HOG and our HOG+C detectors with varying cell sizes. As observed in Figures 4.10b and 4.11b 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.

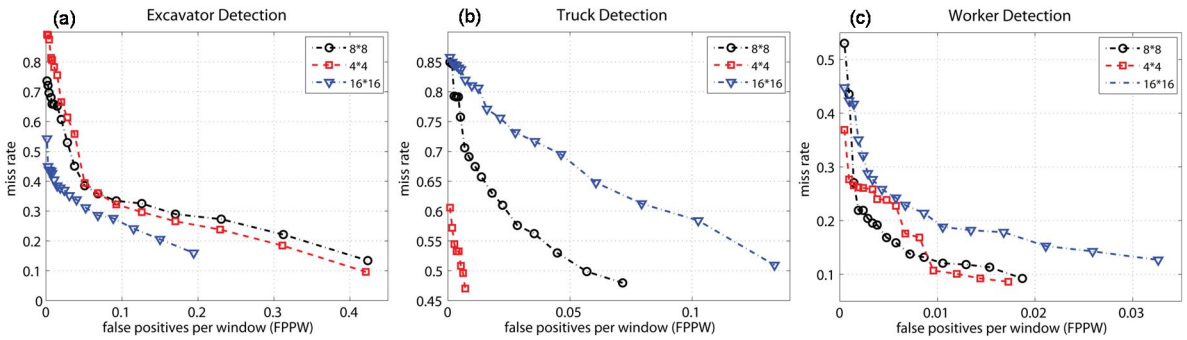

Figure 4.10. Effect of the cell size on performance of HOG for detection of different construction resources.

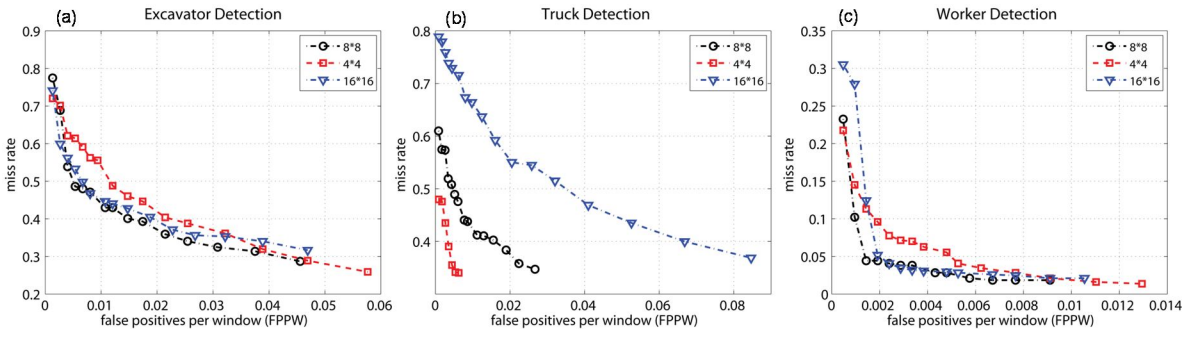

Figure 4.11. Effect of the cell size on performance of HOG+C for detection of different construction resources.

a.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. Figure 4.12 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.

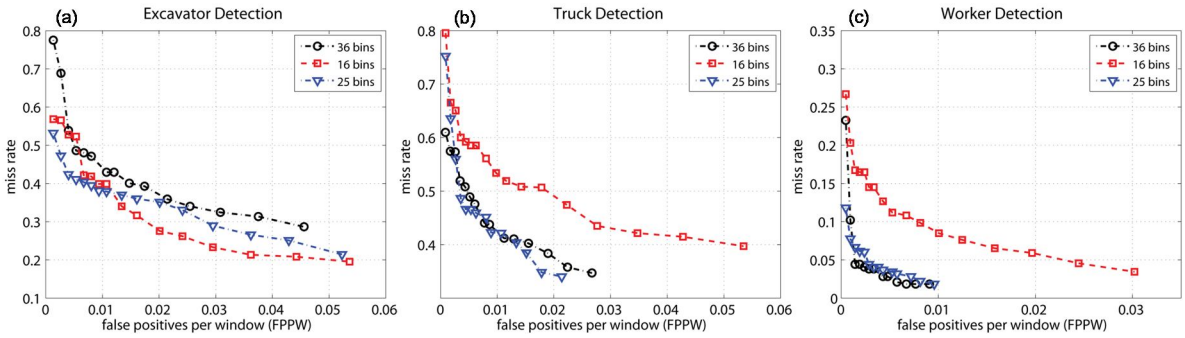

Figure 4.12. Effect of the number of bins in HOC on performance of HOG+C for detecting different construction resources.

4.1.2 HOG+C Deformable Part-based Models

In our proposed HOG+C deformable part-based method (HOG+C-DPM), the detectors have the following properties:

- Linear gradient [-1; 0; 1] voting into 9 orientation bins in 0-180˚ regard the contrast insensitive features and 18 orientation bins in 0-360˚ regard contrast sensitive features are used for generating all HOG descriptors (Felzenszwalb et al. 2010);

- L2-normalized blocks with 4 cells containing 8×8 pixels were used to generate HOG descriptors for excavators, dump trucks, and workers;

4.1.2.1 Implementation Details

In this section the procedure for initializing the structure of the model is described. The LSVM optimization algorithm is susceptible to local minima and thus sensitive to initialization (Felzenszwalb et al. 2010). This initialization and training process contains three phases,

Phase 1- Initializing Root Filters

For training a model with m components, we sort the bounding boxes in P by their aspect ratio and split them into m groups of equal size $P_1, \ldots, P_m$. Aspect ratio is used as a simple indicator of extreme intra-class variation (e.g., in term of excavator, the appearance from side view is completely different than a front view). Next, m different root filters $F_1, \ldots, F_m$, are trained for each group of positive bounding boxes using classical SVM.

Phase 2- Merging Components

Then, the initial root filters are combined into a mixture model with no parts.

Phase 3- Initializing Part Filters

The initialization of part filters follows a simple heuristic. The number of parts is fixed to eight; using a small pool of rectangular part shapes we greedily place parts to cover high-energy regions of the root filter. Once a part is placed, the energy of the covered portion of the root filter is set to zero, and then we look for the next highest-energy region, until eight parts are chosen. The part filters are initialized by interpolating the root filter to twice the spatial resolution.

Figure 4.13-4.15 shows deformable part-based models which are learned for root filters, part filters, and deformation weights using the excavator, truck, and worker training dataset, respectively.

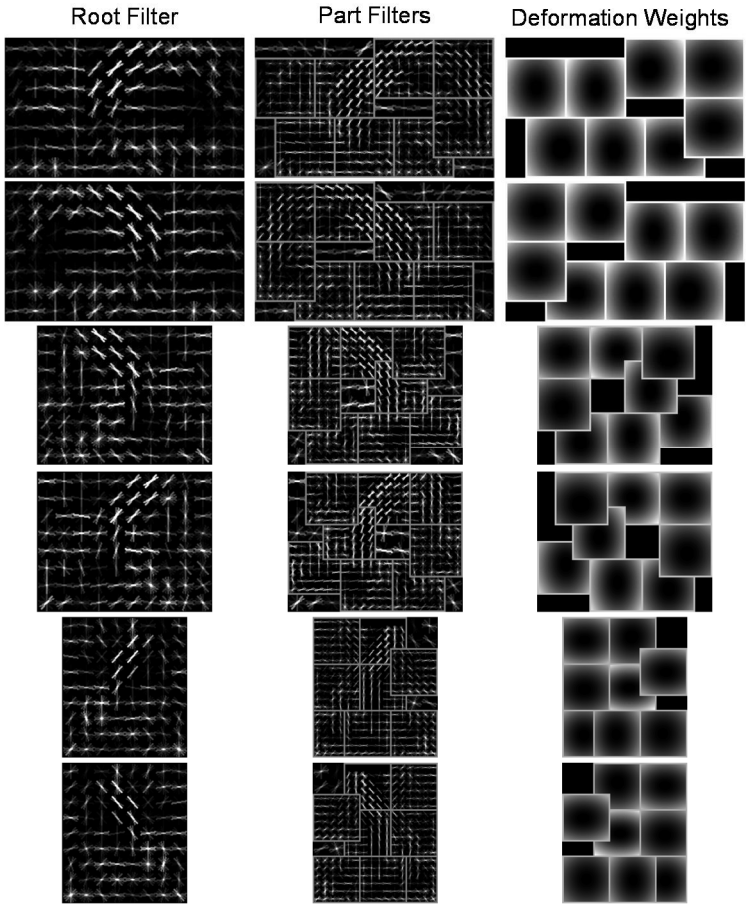

Figure 4.13. Deformable part-based models for excavator dataset: The columns show the root filter, part filters, and deformation weights, respectively.

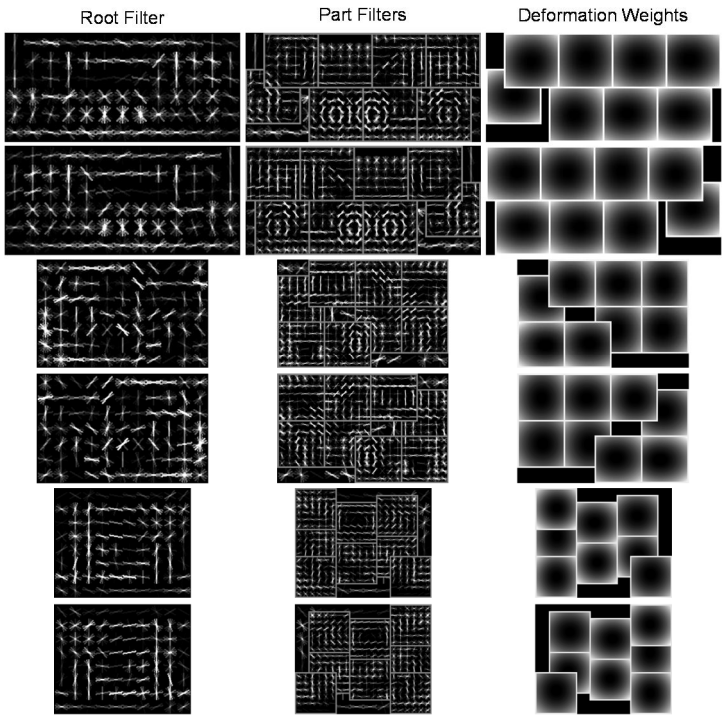

Figure 4.14. Deformable part-based models for truck dataset: The columns show the root filter, part filters, and deformation weights, respectively.

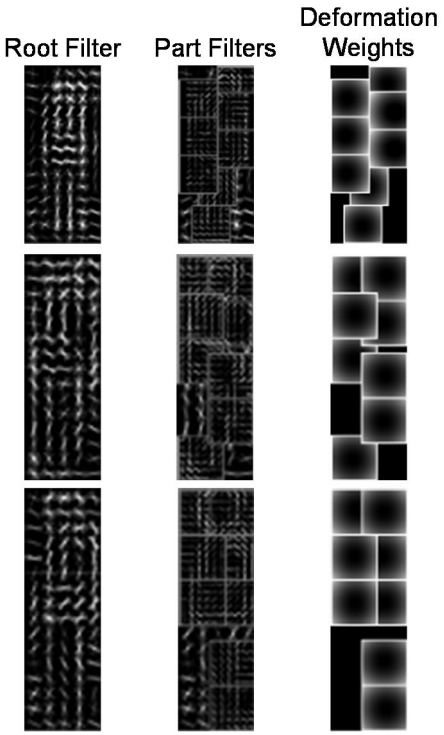

Figure 4.15. Deformable part-based models for worker dataset: The columns show the root filter, part filters, and deformation weights, respectively.

In our testing phase, the detection window slides at 55 scales. 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. For validating the performance of HOG+C-DPM compared to HOG+C sliding detection window we have used more challenging database for worker class. Figure 4.16 shows the DET and precision-recall curves for HOG+C-DPM, HOG+C, and HOG detectors and compares their performances for all three categories of resources. As observed, the new method based on HOG+C-DPM significantly improves the performance of detecting construction resources while eliminate false positives which is really important in detection of resources on noisy and dynamic construction jobsites. In particular it achieves lower miss rates in lower FPPWs and also higher precisions in higher recall values.

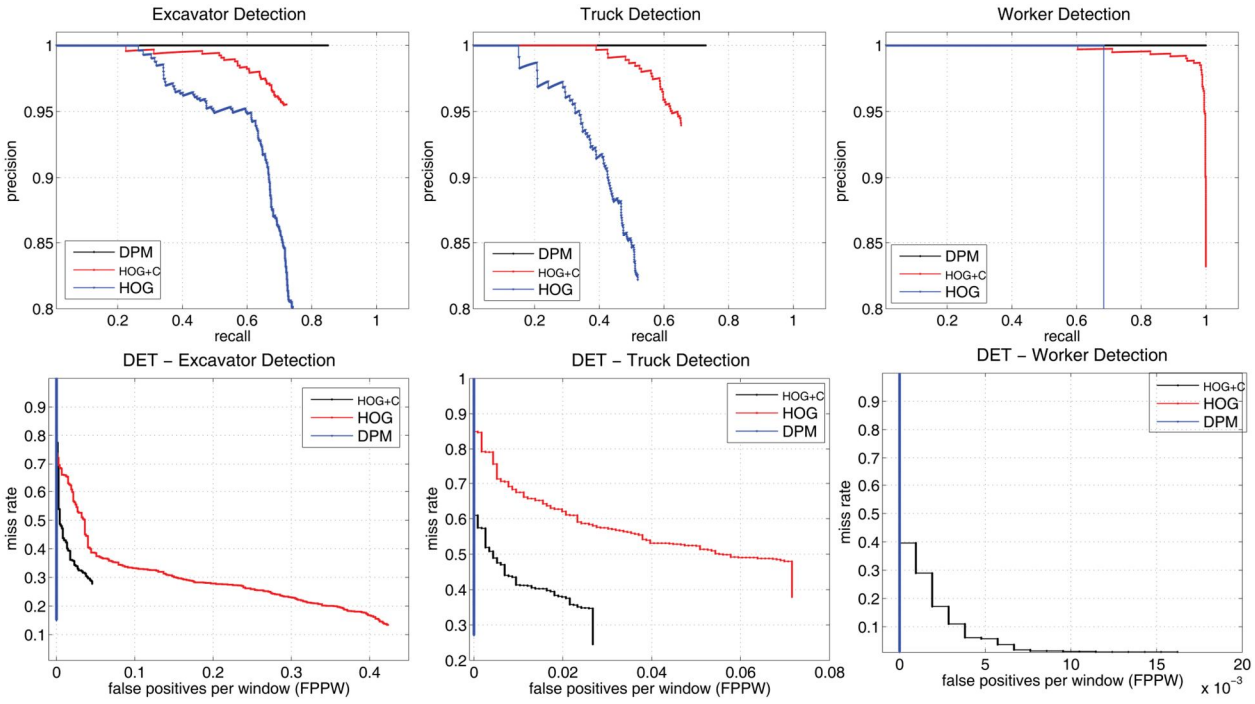

Figure 4.16. Overall results on performance of HOG, HOG+C, and proposed HOG+C-DPM (DPM in the figure) on detection of construction resources. (a-c) precision-recall and (d-f) DET curves for detection of excavators, trucks, and workers, respectively.

The average accuracies in detection of each resource are listed in Table 4.2. 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 these experiments, we have view-independent models for excavators and trucks; i.e., all possible viewpoints are considered together. As a result, our HOG+C-only classifiers result in lower accuracies. Nevertheless, due to the distinct colors of equipment, adding the parts information and forming HOG+C deformable part-based models significantly improves their performance.

Table 4.2. Average accuracies for detection of different construction resources (%)

Resources	**HOG**	**HOG+C**	**HOG+C DPM**
Worker	89.32	93.18	100
Excavator	74.28	82.10	92.02
Truck	76.92	84.88	89.69

4.1.2.2 Discussion on Separate Models for Different Viewpoints

In the following subsections, we systematically study the effects of learning separate models for different viewpoints. For this purpose, we divided the excavator and truck datasets to two different side-viewpoint and front-rear-viewpoint datasets and evaluated the performance of the proposed HOG+C-DPM method in comparison to HOG+C sliding detection window on these datasets. In the case of workers, the body posture and configurations do not change significantly from different viewpoints. It should be noticed that separating the models for different viewpoints increase the computation time, because it needs to learn two models per resource.

Figures 4.17 and 4.18 show the separate models which are learned for different viewpoints of excavator and truck datasets, respectively. We evaluated the performance of HOG+C-DPM compared to HOG+C sliding detection window algorithm on these datasets.

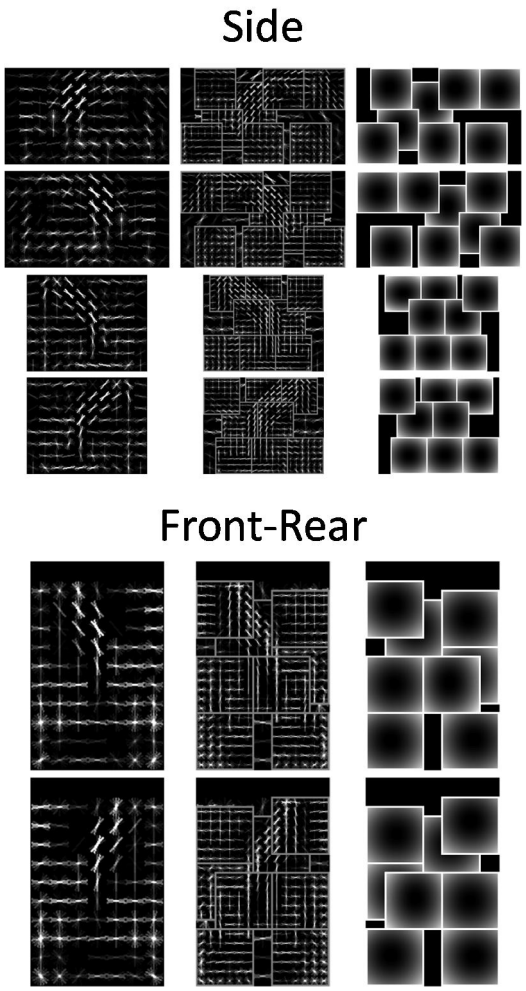

Figure 4.17. The side and front-rear viewpoint models for learned for excavator dataset.

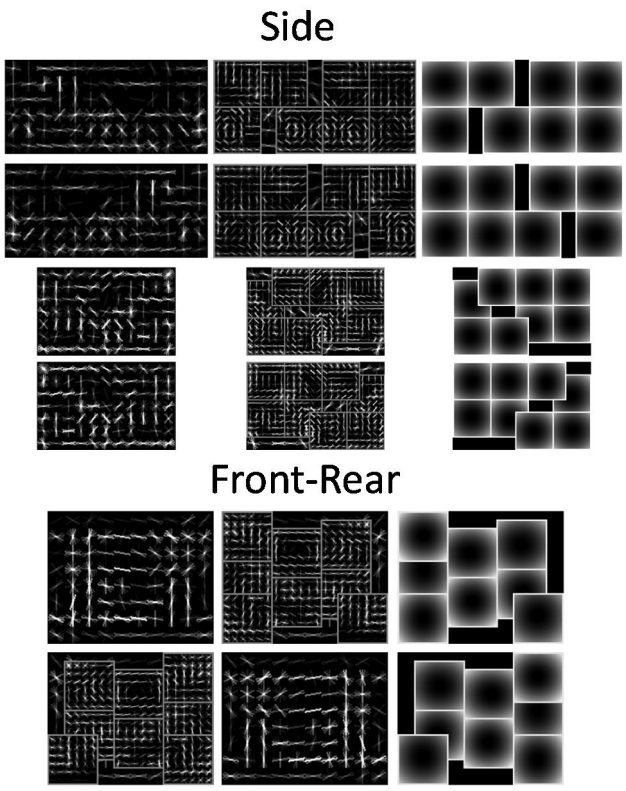

Figure 4.18. The side and front-rear viewpoint models for learned for truck dataset.

Figure 4.19 shows the precision and recall trade-off for the performance of these methods on detection of side-view and front-rear-view excavators and trucks datasets. Table 4.3 summarizes the average accuracies of detection performance on these datasets. As it is clear, the side-view model for excavators improves the detection accuracy compared to the general model while the front-rear-view model decreases the detection accuracy. This is due to lack of enough data from the front and rear view compared to side views of excavators. In the case of truck dataset, the front-rear-view model increases the detection accuracy while the side-view model does not improve the detection accuracy. It can be concluded from these experiments that, by using the robust and powerful HOG+C Deformable Part-based Models, learning separate models for different viewpoints have no significant effect on improving the detection performance while

in some cases it can decrease it. One the main components of HOG+C-DPM method is that it learn multiple models per category based on the aspect ratio of bounding boxes which alleviate the effect of intra-class variability and there is no need for dividing the datasets.

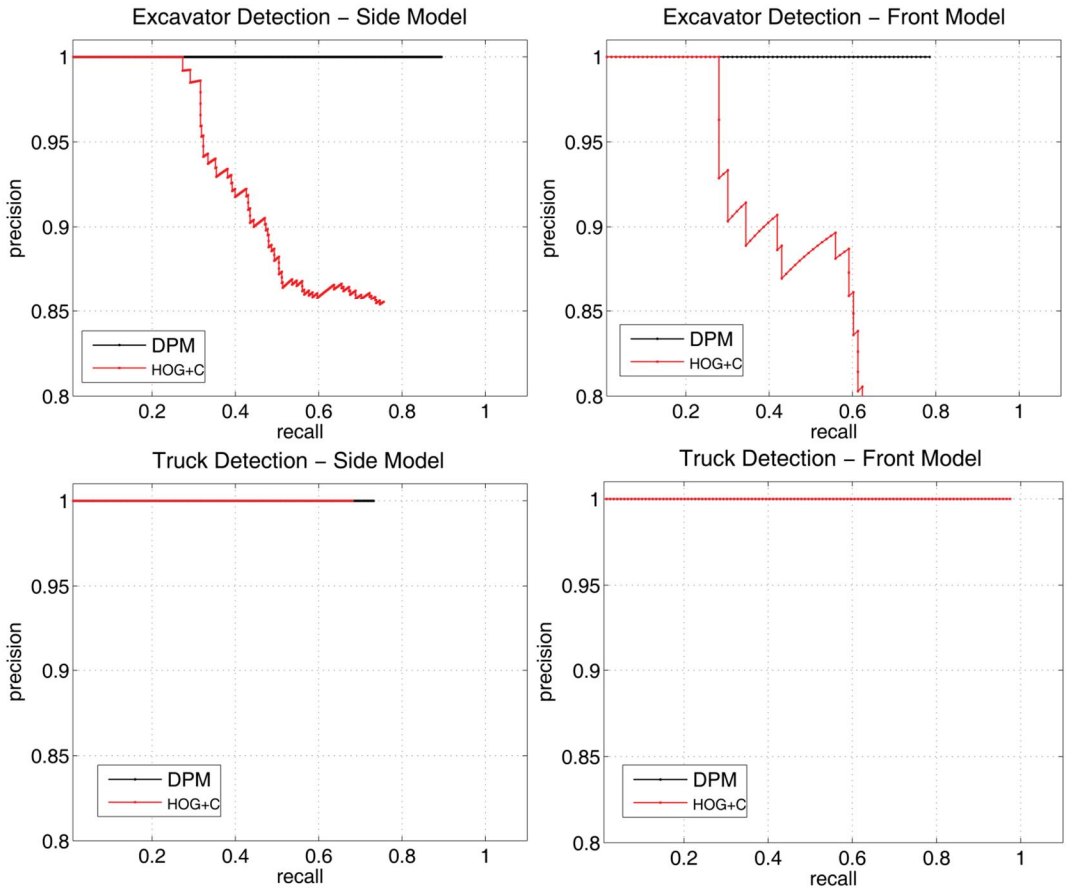

Figure 4.19. precision-recall trade-offs for the performance of separate side and front-rear view models: (a-b) Excavator side and front-rear (front in the figure) models' performance, (c-d) Ttruck side and front-rear (front in the figure) models' performance.

Table 4.3. Average accuracies of learning separate models for detection of different viewpoints (%)

Resources	HOG+C	HOG+C DPM
Excavator General	82.10	92.02
Excavator Side	82.11	94.94
Excavator Front	74.61	89.64
Truck General	84.88	89.69
Truck Side	84.72	87.15
Truck Front	98.88	94.42

4.2 Multiple Resource Detection

In the previous section, we evaluated the performance of our part-based detection method on isolated video frames in which the expectation was to detect a single resource. Here we focus on evaluation of the proposed method for detection of multiple resources with varying degrees of occlusion. The ability of analyzing multi-scale 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.

4.2.1 HOG+C Sliding Detection Window Technique

Figure 4.20 shows the impact of different level of window overlap on accuracy of detecting and localizing 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.

Figure 4.20. Effect of the detection window overlap in accuracy of localizing construction resources in 2D: (a) without overlap, (b) 50% overlap, (c) 98% overlap.

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. Figure 4.21 shows the performance of our algorithm for detection of an excavator in a video sequence where the pose of the equipment was rapidly changing. For example, in Figure 4.21 (a) and (b) the scale of detection window is different regard the different pose of the equipment.

Figure 4.21. Detecting an excavator in a video sequence where in the pose is rapidly changing.

Figures 4.22, 4.23, and 4.24, show the performance of our detector window in detecting multiple resources. As illustrated in Figure 4.22, the 90% overlap, enables excavators and trucks that are working in close proximity to each other to be robustly detected. Figure 4.23 shows another example on detection of construction crew working in proximity to an excavator. This is a critical component for safety assessment purposes. Figure 4.24 shows the performance of our algorithm in detection of multiple excavators in different distance from the camera (scale) and from multiple viewpoints. Full demoes can be found at http://www.raaamc.cee.vt.edu/hogc.

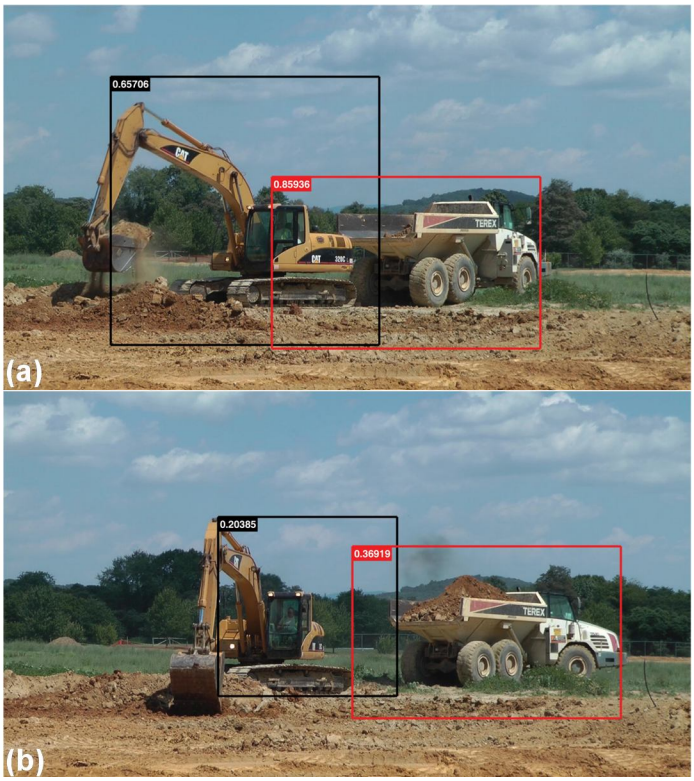

Figure 4.22. Detection in a video sequence where in an excavator and a truck are working in the proximity of each other.

Figure 4.23. Detection of excavators and construction workers in proximity of each other.

Figure 4.24. Example of the capability of our proposed method in detection of multiple excavators with different viewpoints and distances to the camera.

4.2.2 HOG+C Deformable Part-based Model

Figure 4.25 shows the performance of HOG+C-DPM on detection of multiple resources in High-Definition (HD) image. As it can be seen in Figure 4.25a, there are three excavators with different models, distances from the camera, pose, and viewpoints that are working in proximity of each other and the proposed method was able to accurately and efficiently detect and localize all of them without false positive. Figure 4.25b shows the performance of the proposed method on detection of standing workers using mobile camera mounted on a hardhat. It should be noticed that while the mobile camera has high amount of distortion, it was able to detect the standing workers.

Figure 4.25. Performance of HOG+C-DPM on HD images and mobile cameras: (a) detection of multiple excavators working in the noisy and dynamic construction jobsite, (b) detection of standing workers by mobile cameras (high amount of distortion).

As mentioned before, 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. Figures 4.26 and 4.27 shows several snapshots of the performance of our algorithm for detection of an excavator and truck

working in proximity of each other, and the standing workers working on the construction jobsite. The full demo can be found at http://www.raamac.cee.vt.edu/hogcdpm.

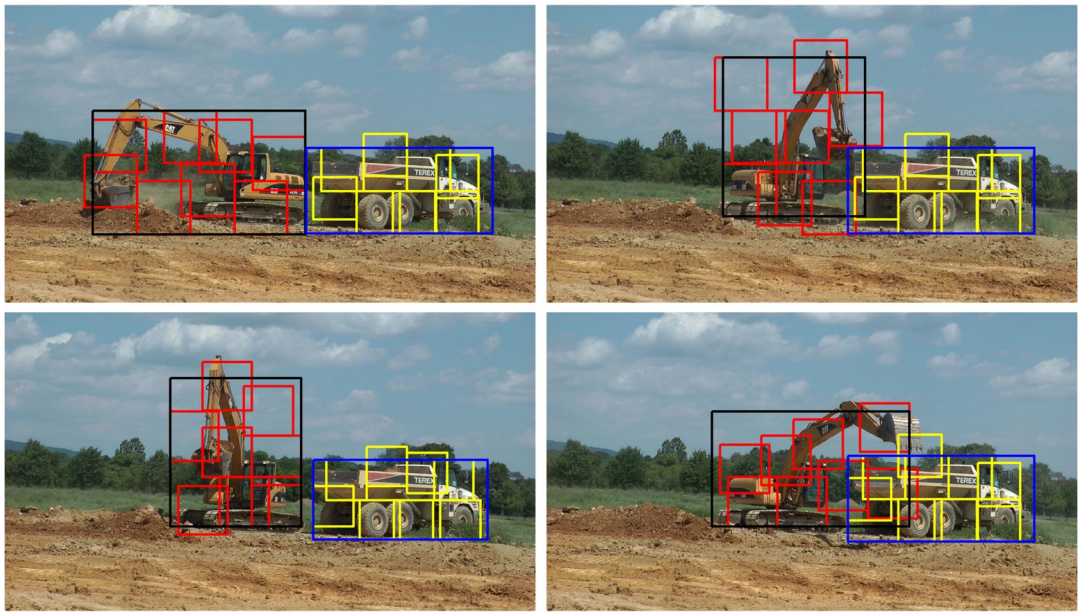

Figure 4.26. Detection of an excavator and a truck working in proximity of each other in video sequences. Full demo can be found at: http://www.raamac.cee.vt.edu/hogcdpm.

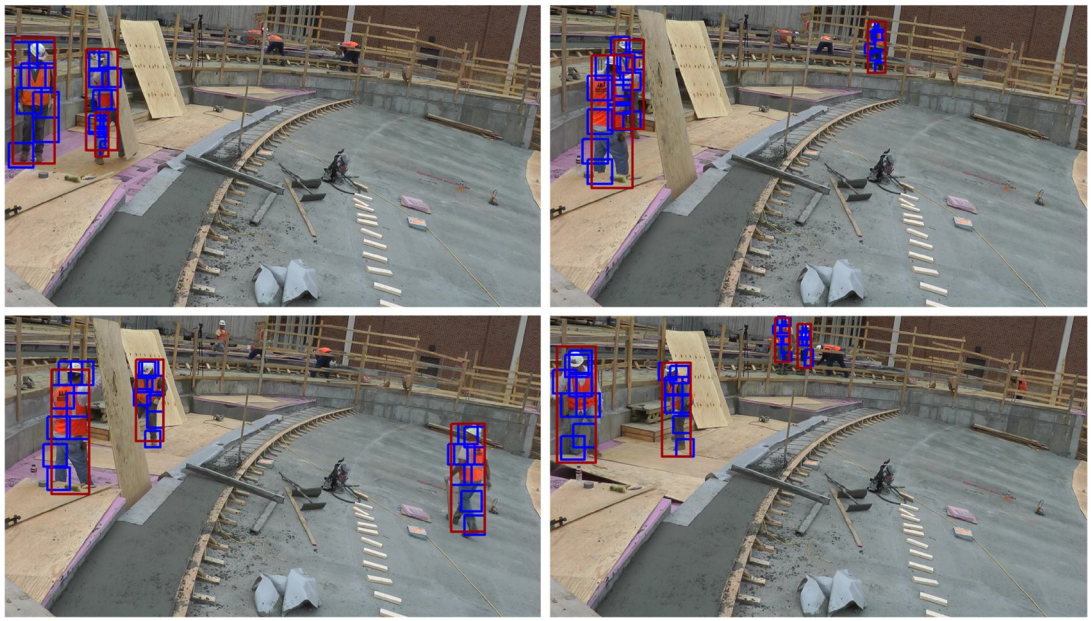

Figure 4.27. Detection of construction workers working in proximity of each other in video sequences. As it can be seen, the proposed method is capable of detection the occluded workers which is really common in dynamic construction sites. Full demo can be found at: http://www.raamac.cee.vt.edu/hogcdpm.

Chapter 5
Conclusions and Future Work

5.1 Discussion on the Proposed Methods 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% in using HOG+C sliding detection window technique and 100%, 92.02%, and 89.69% using HOG+C deformable part-based models, respectively. The ability to detect idling resources, distinguishes our work from 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 $(80 \sim 800) \times (80 \sim 800)$ and $(100 \sim 800) \times (100 \sim 800)$ pixels per excavator and dump truck, and $(50 \sim 700) \times (50 \sim 700)$ pixels per worker, promises the applicability of the proposed method for existing site video cameras. The HOG+C sliding detection window technique which could be real-time is appropriate for construction workers' safety analysis while HOG+C deformable part-based model which provide information about different parts of the object is more appropriate for action recognition and pose estimation purposes. While this thesis 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:

- **Real-time 2D detection and localization in long video sequences.** The presented algorithms are 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 methods 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.

- **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.

5.2 Contributions

As a step towards fully automated performance assessment methods, this research presents two new computer vision based algorithms (HOG+C Sliding Detection Window Technique, and HOG+C Deformable Part-based Models) for automated 2D detection and localization of

construction resources from site video streams in support of automated performance assessment of construction operations.

Moreover, a new comprehensive benchmark dataset (First comprehensive dataset in AEC community) containing over 8,000 annotated video frames including equipment (e.g., excavator and truck) and workers from different construction projects is introduced. Due to lack of existing datasets for benchmarking visual detection construction workers and equipment, it was necessary to create a new comprehensive dataset. This dataset contains a large range of pose, scale, background, illumination, and occlusion variation.

5.3 Practical Significance

The outcome of proposed approach could provide visual information about the various resources working on the jobsite for further analysis such as 3D tracking, action recognition, safety analysis, and performance assessment. 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. Moreover, compared to other sensing technologies (e.g., RFID, GPS, Ultra Wind Band), application of the proposed methods seems more applicable and practical, as it does not require "tagging" of construction entities. Furthermore application of sliding detection window in these proposed methods provides two key benefits:

- Detection of workers and equipment while idle; i.e., it examines static windows for possible resource candidates and is not limited to detection of moving foreground objects (typical in background subtraction techniques).

- 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 assessment.

By implementing the HOG+C sliding detection window technique using NVIDIA CUDA parallel computation framework, the real-time performance can be achieved. Integration of real-time 2D detection and 3D tracking and localization is key components for safety analysis of workers working in proximity of equipment. HOG+C deformable part-based model is a pre-requisite for integrated activity analysis of construction operations which enables project managers to:

- Study their operation automatically
- Revise their construction plan and operation strategies
- Simultaneously reduce their environmental impacts and increase/maintain the level of productivity.

5.4 Conclusions

In this research, we presented two novel methods for automated 2D detection and localization of construction workers (standing) and equipment from site video streams based on using histograms of oriented gradients and Hue-Saturation colors. Our results with average performance accuracies of 100%, 92.02%, and 89.69% for workers, excavators, and dump trucks respectively, hold the promise of applicability of the proposed methods for first step of automated performance assessments. As validated, adding histogram of Hue-Saturation colors to oriented gradients significantly improved the detection of resources. We also evaluated the effect

of different model parameters on detection accuracy. The proposed approaches are independent to scale and viewpoint of resources, as well as illumination conditions. The HOG+C sliding detection window technique which could be real-time is appropriate for construction workers' safety analysis while HOG+C deformable part-based model which provide information about different parts of the object is more appropriate for action recognition and pose estimation purposes. Despite the good performance of proposed HOG+C-based methods, 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. The future work involves implementing the HOG+C based sliding detection window algorithm using the NVIDIA CUDA parallel computing framework which 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, and capability of tracking detected objects are another open research questions in this area.

References

Azar, E.R., McCabe, B. (2011). "Automated visual recognition of dump trucks in construction videos" *ASCE Journal of Computing in Civil Engineering*, In Press.

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

Bay, H., Ess, A., Tuytelaars, T., van Gool, L. (2008). "SURF: Speeded Up Robust Features" *Computer Vision and Image Understanding*, 110(3), 346-359.

Brilakis, I., Park, M., Jog, G. (2011). "Automated vision tracking of project related entities" *Advanced Engineering Informatics*, 25, 713-724.

Caldas, C.H., Grau, D., Haas, C.T. (2006). "Using Global Positioning System to Improve Materials-Locating Processes on Industrial Projects" *ASCE Journal of Construction Engineering and Management*, 132(7), 741-750.

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

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

Christopher, J.C., Burges, A. (1998). "A tutorial on support vector machines for pattern recognition" *Data Mining and Knowledge Discovery*, 2, 121-167.

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

Dalal, N., Triggs, B. (2005). "Histograms of Oriented Gradients for Human Detection." *Proc. IEEE CVPR*, San Diego, CA, 2, 886-893.

Dalal, N., Triggs, B., Schmid, C. (2006). "Human detection using oriented histograms of flow and appearance" *Proc. ECCV* 2, 428-441.

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

ENR (2011). "Don't Blame The Workers" *Engineering News-Record*.

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

Felzenszwalb, P.F., McAllester, D., Ramanan, D. (2008). "A discriminatively trained, multiscale, deformable part model" *Proc. of IEEE CVPR*.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D. (2010). "Object detection with discriminatively trained part-based models" *Proc. of TPAMI*, 32(9), 1627-1645.

Fontana, R.J., Gunderson, S.J. (2002). "Ultra-Wideband Precision Asset Location System" *IEEE Conference on Ultra Wideband Systems and Technologies*, Baltimore, MD.

Forsyth, D., Ponce, J. (2011). "Computer Vision: A Modern Approach" *Second Edition, Pearson Education Inc.*

Golparvar-Fard, M., Pena-Mora, F., Arboleda, C.A., Lee, S. (2009). "Visualization of Construction Progress Monitoring with 4D Simulation Model Overlaid on Time-Lapsed Photographs" *ASCE Journal of Computing in Civil Engineering*, 23(6), 391-404.

Golparvar-Fard, M., Pena-Mora, F., Savarese, S. (2009). "Application of D4AR – A 4-Dimensional augmented reality model for automating construction progress monitoring data collection, processing and communication" *ITcon*, 14, 129-153.

Golparvar-Fard, M., Pena-Mora, F., Savarese, S. (2011). "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), 1099-1116.

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

Gong, J., Caldas, C.H. (2009). "Construction site vision workbench: A software framework for real-time process analysis of cyclic construction operations" *Proc. ASCE Int. Workshop on Computing in Civil Engineering*, Auton, TX, 64-73.

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

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

Goodrum, P.M., Haas, C.T., Caldas, C.H., Zhai, D., Yeiser, J., Homm, D. (2011). "Model to Predict the Impact of a Technology on Construction Productivity" *ASCE Journal of Construction Engineering and Management*, 137.

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

Heydarian, A. (2011). "Vision-based Tracking and Action Recognition of Earthmoving Construction Operations for Automated Productivity and Carbon Footprint Assessments" *M.Sc. Thesis*, Vecellio Construction Engineering and Management, Via Department of Civil and Environmental Engineering, Virginia Tech.

Kamat, V.R., Akula, M. (2011). "Integration of Global Positioning System and Inertial Navigation for Ubiquitous Context-Aware Engineering Applications" *Proc. National Science Foundation Grantee Conference*, Atlanta, GA.

Laptev, I. (2006). "Improvements of object detection using boosted histograms" *BMVC*.

Lowe, D.G. (2004). "Distinctive Image Features from Scale-Invariant Keypoints" *International Journal of Computer Vision*, 60(2), 91-110.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M. (1997). "The DET Curve in Assessment of Detection Task Performance" *NIST*.

Memarzadeh, M., Golparvar-Fard, M., Niebles, J.C. (2012a). "Automated 2D Detection of Consturction Equipment and Workers from Site Video Streams Using Histograms of Oriented Gradients and Colors: *Automation in Construction*, In Press.

Memarzadeh, M., Heydarian, A., Golparvar-Fard, M., Niebles, J.C. (2012b). "Real-time and automated recognition and 2D tracking of construction workers and equipment from site video streams" *ASCE Int. Workshop on Computing in Civil Engineering*, Florida, 429-436.

Navon, R. (2005). "Automated project performance control (appc) of construction projects" *Automation in Construction*, 14(4), 467-476.

Navon, R., Sacks, R. (2007). "Assessing research in automated project performance control (APPC)" *Automation in Construction*, 16(4), 474-484.

NIST (2011-2012). "Criteria for performance excellence" *National Institute of Science and Technology*.

Oglesby, C.H., Parker, H.W., Howell, G.A. (1989). "Productivity Improvement in Construction" *McGraw-Hill*, New York, NY.

Park, M., Koch, C., Brilakis, I. (2011). "3D Tracking of Construction Resources Using an On-site Camera System" *ASCE Journal of Computing in Civil Engineering*, In Press.

Pepik, B., Stark, M., Gehler, P., Schiele, B. (2012). "Teaching 3D Geometry to Deformable Part Models" *Proc. of IEEE CVPR*.

Song, J., Caldas, C.H., Ergen, E., Haas, C.T., Akinci, B. (2004). "Field Trials of RFID Technology for Tracking Pre-Fabricated Pipe Spools" *Proc. of the 21st International Symposium on Automation and Robotics in Construction*.

Song, J., Haas, C.T., Caldas, C.H. (2006). “Tracking the location of materials on construction job sites” *ASCE Journal of Construction Engineering and Management*, 132(9), 911-918.

Su, Y., Liu, L. (2007). “Real-time construction operation tracking from resource positions” *ASCE Int. Workshop on Computing in Civil Engineering*, Pittsburg, PA, 200-207.

Sun, M., Savarese, S. (2011). “Articulated part-based model for joint object detection and pose estimation” *Proc. of ICCV*.

Teizer, J., Lao, D., Sofer, M. (2007). “Rapid automated monitoring of construction site activities using ultra-wideband” *Proc. of 24th Int. ISARC*, Kerala, India, 23-28.

Van de Weijer, J., Schmid, C. (2006). “Coloring local feature extraction” *Proc. ECCV*, 2, 332-348.

Viola, P., Jones, M. (2001). “Rapid object detection using boosted cascade of simple features” *IEEE CVPR*, Kauai, HI, 1, 1-9.

Yang, J., Vela, P.A., Teizer, J., Shi, Z.K. (2011). “Vision-Based crane tracking for understanding construction activity” *Proc. ASCE Int. Workshop on Computing in Civil Engineering*, Miami, FL, 258-265.

Yang, Y., Ramanan, D. (2011). “Articulated pose estimation with flexible mixtures-of-parts” *IEEE CVPR*, 1385-1392.

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