Automated Vision-Based Tracking and Action Recognition of Earthmoving Construction Operations

Arsalan Heydarian

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 Engineering

Mani Golparvar-Fard, Committee Chair
Jesus M. de la Garza
Linsey C. Marr
Juan Carlos Niebles (Universidad del Norte)

April 30, 2012
Blacksburg, VA

Keywords: Construction Performance Monitoring, 2D Tracking, Action Recognition, Histogram of Gradients, Support Vector Machine

Automated Vision-Based Tracking and Action Recognition of Earthmoving Construction Operations

Arsalan Heydarian

Abstract

The current practice of construction productivity and emission monitoring is performed by either manual stopwatch studies which are significantly labor intensive and subject to human errors, or by the use of RFID and GPS tracking devices which may be costly and impractical. To address these limitations, a novel computer vision based method for automated 2D tracking, 3D localization, and action recognition of construction equipment from different camera viewpoints is presented. In the proposed method, a new algorithm based on Histograms of Oriented Gradients and hue-saturation Colors (HOG+C) is used for 2D tracking of the earthmoving equipment. Once the equipment is detected, using a Direct Linear Transformation followed by a non-linear optimization, their positions are localized in 3D. In order to automatically analyze the performance of these operations, a new algorithm to recognize actions of the equipment is developed. First, a video is represented as a collection of spatio-temporal features by extracting space-time interest points and describing each with a Histogram of Oriented Gradients (HOG). The algorithm automatically learns the distributions of these features by clustering their HOG descriptors. Equipment action categories are then learned using a multi-class binary Support Vector Machine (SVM) classifier. Given a novel video sequence, the proposed method recognizes and localizes equipment actions. The proposed method has been exhaustively tested on 859 videos from earthmoving operations. Experimental results with an average accuracy of 86.33% and 98.33% for excavator and truck action recognition respectively, reflect the promise of the proposed method for automated performance monitoring.

Acknowledgements

I would like to express my gratitude to Professor Mani Golparvar-Fard for his supervision and invaluable advice. With his enthusiasm, inspiration, and great efforts to explain things clearly and simply he helped me to achieve beyond the expectations of a Master's student. He has gone above and beyond the normal expectations of an advisor, and I greatly attribute much of my success to him. He invigorates a lot of passion and motivation in his students and has definitely contributed to my future decisions as pursuing a Ph.D. degree.

Special thanks to Professor de la Garza for motivating me to pursue my Master's degree at Virginia Tech and his continuous support during my undergraduate and graduate studies, and to this date.

I would also like to express my gratefulness to my parents, Siamak Heydarian, and Niloufar Yashmi, as well as my brother Nima Heydarian, for creating this opportunity for my higher education achievements and their continuous support throughout my Master's program during happy and stressful periods.

I would like to thank my friend and colleague, Milad Memarzadeh, for his collaboration and support on some parts of this research, in which without his help this work would have not been as complete and detailed.

Finally, this research would have not been possible without the support of Virginia Tech Department of Planning, Design, and Construction, Holder, and Skanska construction companies. The support of current and former RAAMAC lab members, in particular Vahid Balali, Chris Bowling, David Cline, Hooman Rouhi, Hesham Barazi, Daniel Vaca, Marty Johnson, Nour Dabboussi, Fabian Capra, Rafael Suriel, and Moshe Zelkowicz is also appreciated.

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

1.1 Research Overview

This thesis provides an overview of the proposed framework shown in Figure 1.1 to detect, spatially locate, and evaluate actions of construction equipment for the purpose of performance assessment of construction operations.

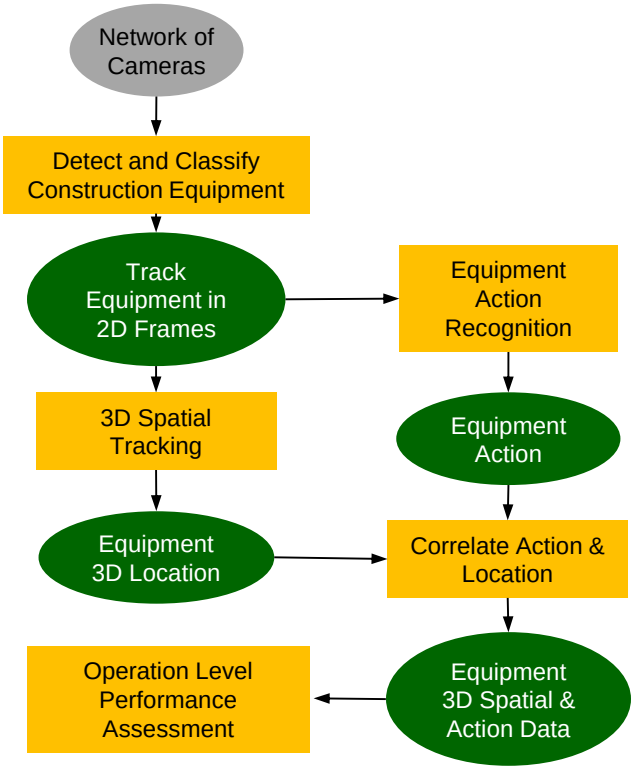


Figure 1.1: Proposed Research Framework

In this research, a network of high resolution video cameras is installed around a construction site which collects operations' video streams and then are transferred to a powerful computer nearby for analysis wherein the construction equipment are recognized in 2D frames. The process of 2D tracking involves traversing video frames using a sliding detection window technique. In the proposed algorithm, the distribution of intensity and hue-saturation colors is formed as Histograms of Gradients and Colors (HOG+C) over sliding detection windows and consequently the equipment categories (e.g., excavator, truck) are automatically recognized through a Multiple Binary Support

Vector Machine (SVM) classifier. These video frames are further processed across the network of cameras to spatially locate the equipment in 3D. In this algorithm, using a triangulation technique based on Direct Linear Transform (DLT) followed by a non-linear optimization, the Epipolar geometry of the detected windows (geometry of stereo vision) is formed and the position of the detected equipment is localized in the site 3D coordinate system. Based on a novel supervised learning method, video streams are further divided into a collection of spatio-temporal features by extracting space-time interest point and hence equipment action categories are recognized. In this algorithm, probability distributions of the spatio-temporal features and the intermediate cluster centers corresponding to equipment action categories are automatically learned using a multi-class binary Support Vector Machine (SVM) classifier. The learned model is further used for categorization and localization of equipment actions (e.g., digging, hauling, dumping, moving, and idle for excavators) in a novel video. These algorithms are the first automated techniques to detect, 3D locate, and recognize actions of construction equipment. The results of this work can facilitate decision-making process on corrective control actions and ultimately minimize construction carbon footprint, while maintaining or increasing productivity through removing equipment idle time, maintenance, properly sizing of equipment, changing sequence of operations, or changing sources of material (not in the scope of this work).

1.2 Research Motivation

Motivations behind the proposed research framework lay in (1) the transformative potential of gradually and inexpensively sensing action and location of construction equipment through a network of cameras installed on a construction site; (2) deficiencies of the current monitoring practices; and (3) the pressing need for reducing emissions and carbon footprint of construction operations;

1.2.1. Automated Sensing Actions and Locations

Over the past few years, cheap and high resolution digital cameras, extensive data storage capacities, in addition to availability of internet on construction sites, have enabled

capturing and sharing of construction image collections and video streams on a truly massive scale. This imagery is enabling construction firms to remotely and easily analyze progress, safety, quality, and productivity (Heydarian and Golparvar-Fard 2011, Golparvar-Fard et al. 2010). Using a network of these high definition video cameras, this thesis proposes a new approach for automated detection of construction location and action which can eventually facilitate remote tracking of construction operations productivity and carbon footprint in future.

1.2.2. Monitoring Deficiencies

Equipment activity analysis, the continuous and detailed process of benchmarking, monitoring, and improving the amount of time construction equipment spends on different construction activities can play an important role in improving construction productivity and minimizing construction carbon footprint. It examines the proportion of time equipment spend on specific construction activities. Combination of detailed assessment and continuous improvement can help minimize the idle time, improve productivity of operations (Gong and Caldas 2011), save time and money (Zou and Kim 2007), and result in reduction of fuel use, construction emissions and carbon footprint (Lewis et al. 2011, EPA 2010). It can also extend equipment engine life and provide safer environment for operators and workers.

Despite the great benefits that activity analysis provides in identifying areas for improvement, implementation, and reassessments, an accurate and detailed assessment of work in progress requires an observer for equipment involved in every construction activity which can be prohibitively expensive. In addition, due to the variability on how construction tasks are carried out, or in the duration of each work step, it is often necessary to record several cycles of operations. Not only are the traditional time-studies labor intensive, but also significant amount of information that needs to be manually collected and analyzed can affect the quality of the process. Furthermore, without a detailed and continuous activity analysis, it is not possible to investigate the relationship between the activity duty cycles versus fuel use and emissions (Frey et al. 2010). There is

a need for a low-cost, reliable, and automated method that can be widely applied across all projects.

1.2.3. Reduction of Operational Emission

A large body of literature has already examined building and infrastructure life cycle assessment and impacts of the greenhouse gas (GHG) emissions generated during operations on the environment (Frey et al. 2010, Shiftehfar et al. 2010, Artenian et al. 2010, Khasreen et al. 2009, Ahn et al. 2009). Nonetheless, the challenges for a global climate change (EPA 2009) is motivating government agencies to investigate strategies on how greenhouse gas (GHG) emissions associated with the construction of buildings and infrastructure could be reduced (Cass and Mukherjee 2010; Santero and Horvath 2009; Peña-Mora et al. 2009; EPA 2009).

The construction industry is considered to be one of the major contributors of these GHG emissions (EPA 2010). According to EPA, historical emission from 14 industrial sectors in the U.S. accounts for 84 percent of the industrial GHG emissions, while the construction sector is responsible for 6 percent of the total U.S. industrial-related GHG emissions, placing the construction sector to be the producer of the third highest GHG emissions along all sectors. The relatively large amount of emission produced in short period of time, reveals the importance of significantly reducing this source of emission. Among all environmental impacts from construction processes (e.g., waste generation, energy consumption, resource depletion, etc.), emissions from construction equipment account for the largest share (more than 50 percent) of the total impact (Skanska 2011, Ahn et al. 2010, Guggemos and Horvath 2006).

In the United States, a new set of EPA off-road diesel emissions regulations is rapidly becoming a concern for the construction industry (ENR 2010) and the controversial issues associated with such regulations have required Associated General Contractors of America and the California Air Resources Board to postpone enforcements of these emission rules until 2014. Although these regulations are

considered to minimize construction carbon footprint by a large factor, yet industry interest has been minimal due to high cost of the alternatives: (1) high cost of new equipment, and (2) upgrading older machinery. These regulations are challenging construction firms to find solutions to reduce the carbon footprint of their operations without affecting productivity and the final cost of their projects. In order to meet these ambitious reductions in carbon footprints, a major cut in GHG emissions due to construction operations, manufacture, and delivery of materials is necessary.

1.3 Research Objectives

This research does not intend to assess all construction operations and track all kinds of construction operations and machinery. Rather the specific focus is on earthmoving construction activities that involve heavy machinery. These operations are always part of critical activities in a construction schedule and have a significant contribution to the carbon footprint of an operation. In addition, the productivity assessment of these operations is also very important as any delay in such critical activities can significantly impact the overall budget and schedule of the project. For this purpose, this research concentrates on detecting, classifying, spatially locating and evaluating actions of construction equipment. The particular focus is on critical and machinery-intensive construction activities that are visible from a network of cameras installed on a job site to automatically monitor their productivity and carbon footprint. The specific research objectives are as follows:

1. Create a comprehensive image dataset for earthmoving operations to facilitate 2D detection of construction equipment;
2. Formalize, create and identify a comprehensive video dataset of all possible types of actions per equipment for automated action recognition;
3. Automatically detect different types of earthmoving equipment from video frames at a reasonably high accuracy;
4. Track 3D location of construction equipment using two or more fixed video cameras; and finally

5. Automatically recognize different actions of construction equipment at reasonable accuracy;

Successful execution of the proposed research will automate several key steps towards a fully automated computer vision based model for construction activity analysis. An automated activity analysis will ultimately transform the way construction operations are currently being monitored. Construction operations will be more frequently assessed through an inexpensive and easy to install solution, thus relieving construction companies from the time-consuming and subjective task of manual method analysis of construction operation, or installation of expensive location tracking and telematics devices.

In the following chapters, first the automated 2D detection and 3D tracking of construction equipment is discussed. Next, the automated action recognition of construction equipment from site video streams is presented. Finally the perceived benefits and the challenges associated with application of the proposed method for the automated tracking and recognition for operational performance improvement is detailed. The conclusions and future work are discussed in chapter 4.

1.4 Manuscript Overview

The thesis consists of an introduction chapter and a conclusion chapter to outline the conducted research. Chapters 2 and 3 present two research papers that focus on specific contributions of this work.

A version of Chapters 2 and three will be submitted to the Elsevier Journal of Automation in Construction and Advanced Engineering Informatics respectively. Particularly in chapter 2, a new algorithm to automatically detect construction equipment in 2D and track in 3D is introduced. The authorship of this chapter is Arsalan Heydarian, Milad Memarzadeh, Juan Carlos Niebles, and Mani Golparvar-Fard.

Chapter 3 will be submitted in Elsevier Journal of Advanced Engineering Informatics. In this paper a new algorithm to automatically recognize actions of the

earthmoving equipment is introduced. The authorship of this chapter is Mani Golparvar-Fard, Arsalan Heydarian, and Juan Carlos Niebles.

Chapter 4 is a conclusion chapter which summarizes the conducted research, the contributions made, and describes the future work.

References

Ahn C., Rekapalli P., Martinez J., and Peña-Mora F. (2009). "Sustainability Analysis of Earthmoving Operations." *Proc., 2009 Winter Simulation Conference*, 2605-2611.

Artenian, A., Sadeghpour, F., and Teizer, J. (2010). "Using a GIS Framework for Reducing GHG Emissions in Concrete Transportation," *Proc., Construction Research Congress*, Canada, May, 1557-1566.

Brilakis, I., Park, M.W. and Jog, G. (2011). "Automated Vision Tracking of Project Related Entities". *Journal of Advanced Engineering Informatics*, Elsevier, 25(4), 713-724.

Bouguet, J.Y. (2011). "Camera Calibration Toolbox for Matlab". http://www.vision.caltech.edu/bouguetj/ (Last accessed May 2011)

Cass, D. and A. Mukherjee. (2010). "Calculation of greenhouse gas emissions associated with highway construction projects using an integrated life cycle assessment approach." *Proc., Construction Research Congress, Banff, Alberta*, 1406-1415.

Dalal, N. and Triggs B. (2005). "Histograms of Oriented Gradients for Human Detection". *Proc., IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2, 886-893.

Dollar, P., Rabuad V., Cottrell G., and Belongie S. (2005). "Behavior Recognition via Sparse Spatio-temporal Features." *2nd joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 65-72.

El-Omari, S., and 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.

U.S. Environmental Protection Agency (EPA) (2010). "Climate Change Indicators in the United States." USEPA #EPA 430-R-10-00.

U.S. Environmental Protection Agency (EPA) (2009). "Potential for Reducing Greenhouse Gas Emissions in the Construction Sector." Sector Strategies.

Golparvar-Fard M., Peña-Mora F. and Savarese S. (2010). "D^4AR – 4 Dimensional augmented reality - tools for automated remote progress tracking and support of decision-enabling tasks in the AEC/FM industry." *Proc., The 6th Int. Conf. on Innovations in AEC*, State College, PA.

Golparvar-Fard M., Peña-Mora F., and Savarese S. (2009). "D^4AR- A 4-Dimensional augmented reality model for automating construction progress data collection, processing and communication." *Journal of Information Technology in Construction (ITcon)*, 14, 129-153.

Gong J., and Caldas C.H. (2010). "Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations." *ASCE J. Comp. in Civ. Engrg. 24*, 252-263.

Grau D. and Caldas C. (2009). "Methodology for Automating the Identification and Localization of Construction Components on Industrial Projects." *ASCE J. Const. Eng. Mgmt*, 23 (1), 3-13.

Guggemos, A. and A. Horvath (2006), "Decision-Support Tool for Assessing the Environmental Effects of Constructing Commercial Buildings," *Journal of Architectural Engineering*, 187-195.

Heydarian A., and Golparvar-Fard M., (2011) "A Visual Monitoring Framework for Integrated Productivity and Carbon Footprint Control of Construction Operations." *ASCE Computing in Civil Eng.*,182(416)62.

Ikizler N., and Forsyth D.A (2008). "Searching for Complex Human Activities with No Visual Examples." *IJCV 80*. 337-357.

Khasreen, M., Banfill, P., and Menzies, G. (2009). "Life-cycle assessment and the environmental impact of buildings: a review". *Sustainability* 1(3), 674–701.

Kockelman K., Bomberg M., Thompson M., and Whitehead C. (2009). "GHG Emissions Control Options - Opportunities for Conservation." National Academy of Sciences.

Laxton B., Lim J., and Kriegman D. (2007). "Leveraging Temporal, Contextual and Ordering Constraints for Recognizing Complex Activities in Video." *CVPR, IEEE*

Laptev I. (2005), "On Space-Time Interest Points." *Int. J. of Computer Vision*, 64, 107-123.

Laptev I., Marszalek M., Schmid C., and Rozenfeld B. (2008). "Learning Realistic Human Actions from Movies." *Proc., Computer Vision and Pattern Recognition*, IEEE, *Conference on Computer Vision and Pattern Recognition.*, 1-8.

Lewis P., Leming M. L., Frey H.C., and Rasdorf W. (2011). "Assessing Effects of Operational Efficiency on Pollutant Emissions of Nonroad Diesel Construction Equipment." *Proc., Transportation Research Board,* 11-3186

Lewis P., Frey H.C., and Rasdorf W. (2009). "Development and Use of Emissions Inventories for Construction Vehicles." *J. of the Transportation Research Board,* 46-53.

Lewis P., Rasdorf W., Frey C., Pang S., and Kim K. (2009). "Requirements and Incentives for reducing Construction Vehicle Emissions and Comparison of Nonroad Diesel Engine Emissions Data Sources." *ASCE J. of Construction Eng. and Mgmt.*, 135 (5), 341-35.

Luers, A. L., M. D. Mastrandrea, K. Hayoe, and P.C. Frumhoff (2007). "How to Avoid Dangerous Climate Change: A Target for U.S. Emissions Reductions." *Union of Concerned Scientists*.

National Research Council (2009). "Committee on Advancing the Competitiveness and Productivity of the U.S. Construction Industry."

Oglesby C.H., Parker H.W., and Howell G.A. (1989). *Productivity Improvement in Construction*. McGraw-Hill, New York, NY, 84-130.

National Institute of Science and Technology (NIST) (2011). "2011-2012 Criteria for Performance Excellence." http://www.nist.gov/baldrige/publications/upload/2011_2012_Business_Nonprofit_Criteria.pdf (accessed September 2011).

Navon R., Goldschmidt E., and Shpatnisky Y. (2004). "A Concept Proving Prototype of Automated Earthmoving Control." *Elsevier J. of Automation in Construction*, 13, 225-239.

Niebles J.C., Wang H., and Fei-Fei Li. (2008). "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words." *International Journal of Computer Vision*, 79(3), 299-318.

Peña–Mora F., Ahn C., Golparvar-Fard M., Hajibabai L., Shiftehfar S., An S., Aziz Z. and Song S.H. (2009). "A Framework for managing emissions during construction." *Proc., Conf. on Sustainable Green Bldg. Design and Construction*, National Science Foundation

Shiftehfar R., Golparvar-Fard M., Peña-Mora F., Karahalios K.G., and Aziz Z. (2010). "The Application of Visualization for Construction Emission Monitoring." *Proc., Construction Research Congress 2010*, Banff, Canada, 1396-1405.

Su Y., and Liu L. (2007). “Real-time Construction Operation Tracking from Resource Positions.” *Proc., ASCE Int. Workshop on Computing in Civil Eng.*, Pittsburgh, PA, 200-207.

Teizer, J. and Vela, P.A. (2009). "Personnel Tracking on Construction Sites using Video Cameras". *Special Issue of the Journal of Advanced Engineering Informatics*, Elsevier, 23(4), 452-462.

U.S. Green Building Council (UCGBC) (2008). “Green Building Facts.” *Environmental Information Administration.*

Yang, J., Arif, O., Vela, P.A., Teizer, J., and Shi, Z. (2010). "Tracking Multiple Workers on Construction Sites using Video Cameras." *Special Issue of the Journal of Advanced Engineering Informatics*, 2(4), 428-434.

Wang Y., and Mori G. (2009). “Human Action Recognition by Semi-latent Topic Models. *IEEE TPAMI* 31. 1762-1774.

Wong S.F., Kim T.K., and Cipolla R. (2007). “Learning Motion Categories Using Both Semantic and Structural Information.” *Proc., Computer Vision and Pattern Recognition*, IEEE, 1-6.

Zou, J., and Kim, H. (2007). “Using Hue, Saturation, and Value Color Space for Hydraulic Excavator Idle Time Analysis.” *ASCE J. Computing in Civil Engineering*, 21, 238-246.

Chapter 2: Automated Video-based Detection and 3D Localization of Multiple Construction Equipment Using HOG+C and Triangulation Methods

2.1 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 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 the companies still need to analyze how to improve. In addition to their immediate benefits, site video streams provide an ideal test bed for developing automated computer vision based performance assessment algorithms that can work effectively 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 requires 1) automated and real-time 2D tracking of the equipment and workers from single cameras; 2) synchronizing detections across multiple cameras and localize the resources in 3D; and finally 3) automated action recognition. Within this scope, one key challenge is automated 2D tracking; i.e., figuring out what resources are visible within a camera's field of view and continuously track them for the entire period of the time the resource is visible. A robust 2D detection provides an opportunity for continuous 3D localization and

action recognition which are critical components for any automated vision-based performance assessment system. While a number of researchers have looked into developing vision-based assessment methods (section 2.2), many challenging open problems remain.

As a step towards fully automated performance assessment methods, this paper focuses on automated 2D detection and 3D localization of construction equipment from onsite video streams. In the proposed framework, a network of fixed high-definition and calibrated cameras is installed around construction sites to record daily construction operations. The video feeds are continuously processed to directly detect frames that contain construction workers and equipment (from now on will be called "resources"). Using low-level features based on Histogram Of Gradients and Hue-Saturation Colors (HOG+C), a new multiple Support Vector Machine (SVM) resource classifier is developed which can recognize and track the dynamic resources in 2D video frames (i.e., worker vs. equipment). Next, using a minimum of two cameras, the detected resources in video frames are processed and are localized in 3D using Direct Linear Transform (DLT) algorithm followed by a non-linear optimization..

2.2 Background and Related Work

The construction industry is still using traditional data collection methods for performance analysis including direct manual observations, stop motion analysis (Oglesby et al. 1989), and survey based methods. Although these methods provide beneficial solutions in terms of improving operation's productivity, yet their implementation is time-consuming, manual and labor-intensive, and can be prone to errors (Gong and Caldas 2011; Zhai et al. 2009). The significant amount of information required to be manually collected may 1) adversely affect the quality of the analysis and make it subjective (Golparvar-Fard et al. 2009; Gong and Caldas 2009; Grau et al. 2009), and 2) minimizes opportunities for continuous benchmarking and monitoring which is a necessary step for performance improvement (NIST 2011). As a result, many critical decisions may be made based on inaccurate or incomplete information, ultimately leading

to project delays and cost overruns. In recent years, several groups have focused on developing techniques that facilitate construction performance assessments. These techniques are categories into sensor-base and vision-based approaches and are as follows:

2.2.1. Current Practice of Sensor Based Tracking

In recent years, a number of research groups have focused on creating and developing techniques that can automatically assess construction performance and facilitate operation idle reductions or improvement of operational efficiency. Gong and Caldas (2011 and 2010), Goodrum et al. (2011), and Su and Liu (2007) all emphasize on the importance of a real-time construction operation tracking of resources for improving construction performance. Different tracking technologies, such as barcodes and RFID tags (Grau et al. 2009; Navon and Sacks 2007; Song et al. 2006; Song et al. 2004), Ultra WideBand (UWB) (Cheng et al. 2011; Williams et al. 2007; Teizer et al. 2007), 3D range imaging cameras (Gong and Caldas 2008; Teizer et al. 2007), global and local positioning systems (GPS) (Grau et al. 2009; Caldas et al. 2006; Ergen et al. 2007) and computer vision techniques, have been applied at construction sites to provide tracking data. Besides their application for material tracking, they have also been used in locating workers in congested or open areas, and recording the sequence of their movement necessary to complete a task. Each one of the technologies has certain shortcomings and advantages that pertain to each application.

UWB technology 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). Teizer et al. (2007) applied the UWB technology for real time material location tracking system; as a result, its ability to provide accurate 3D locations in real-time is a benefit to tracking resources on construction sites. Although this system is promising for tracking, it is not sufficient for construction sites due to the need for carrying satisfactory positioning data to the system prior to the implementation of the UWB system (Brilakis et al. 2011). This technology is

also considered to be costly for construction sites since it requires the installation of sensors on every entity being tracked and as a result cannot be used on workers. Recent researches have tested the use of 3D range imaging camera on construction sites for spatial modeling (Gong and Caldas 2008) and resource tracking (Teizer et al. 2007). However, low resolutions and short range of these cameras make these systems difficult and insufficient to be used on large scale construction sites.

GPS modules have also been applied to construction practices such as positioning of equipment and surveying (Caldas et al. 2006). Despite the wide range of benefits that GPS can offer to the construction industry, using it for indoor tracking of workers will be very limited. GPS can only operate outdoor, and needs to be regularly attached to the resource that is being tracked; therefore, tracking construction resources with GPS is infeasible in many 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 seems to be accurate for tracking workers outdoors, nonetheless, its accuracy degrades with both path complexity and 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. This makes the application of such systems unattractive.

In the case of RFID tags, although they have high durability in harsh environments, do not require line-of-sight, and can be embedded in concrete, yet they are not effective for construction sites. Unless combined with other tools and technologies, RFID can only function within radius inside which the track resource exists (Brilakis et al. 2011). Furthermore, a tag needs to be attached to each resource that is being tracked and due to its near-sighted effects it has limitations for real-time tracking applications. El-Omari and Moselhi (2009); Ergen et al. (2007); Navon and Sacks (2006); and Navon (2005) all introduced different techniques of automated localization and tracking of construction equipment using RFID combined with GPS technology. Despite the potential, RFID tags still require a comprehensive infrastructure to be installed on the jobsite, the near-sightedness of RIFD still limits the applicability of real-time tracking,

and due to GPS applications, the line-of-sight in many locations may adversely impact their benefits. Even in most ideal working scenarios, these technologies can only provide accurate location information. For a comprehensive assessment of performance, worker and equipment action information is required; nonetheless the nature of these technologies does not enable collection of such information.

2.2.2. Current Vision Based 2D Resource Tracking

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 which is a key benefit for activity analysis and formation of crew-balance charts for craft productivity assessment purposes. Despite the popularity of onsite observations 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 still limited to certain projects. Several recent studies such as (Gong et al. 2011; Navon and Sacks 2007; Brilakis et al. 2011; Golparvar-Fard et al. 2009a; Golparvar-Fard et al. 2009b) have emphasized on the need for developing automated video-based techniques. Development of automated video-based methods for action recognition or 3D resource tracking, at first requires the location of the workers and equipment to be tracked in 2D. Recently developed methods such (Gong et al. 2011; Zou and Kim 2007) are either simulated in controlled environments or have primarily focused on automating the 3D tracking assuming semi-automated tracking of resources in 2D. (Brilakis et al. 2011; Park et al. 2011; Yang et al. 2011) use priori knowledge for their assessments such as expected known locations for tracking tower crane (Yang et al. 2011), or application of Scale Invariant Feature Transforms (SIFT) (Lowe 2004) or Speeded Up Robust Features (SURF) (Bay et al. 2008) for initial recognition which can limit their applications in uncontrolled and dynamic conditions.

Two recent works (Chi and Caldas 2011; Rezazadeh-Azar and McCabe 2011) focus on developing techniques for learning, automated 2D tracking, 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, background subtraction 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 in (Rezazadeh-Azar and McCabe 2011) for learning and 2D tracking off-highway dump trucks in video streams. Particularly the application of HOG detectors (Dalal and Triggs 2005), Haar-like detectors (Viola and jones 2001), Haar-HOG cascade (Negri et al. 2008), and Blob-HOG cascade methods are proposed. Due to application of background subtraction, this work is not also able to recognize idle resources.

2.2.3. Current Vision Based 3D Resource Localization

Although recent studies have emphasized on the need for cost effective monitoring techniques, to-date none of the existing methods could simultaneously locate equipment in 3D, and more importantly, recognize their actions. Several researchers including Brilakis et al. (2011) and Gong and Caldas (2011) have proposed vision-based methods for tracking project entities that have potential in addressing some of these requirements. However, current vision-based methods (e.g., Gong and Caldas 2011, Zou and Kim 2007) are either simulated in controlled environments or they have only looked into automating one component of the overall method (mostly 3D tracking). Other vision-based location tracking approaches (e.g., Yang et al. 2011, Brilakis et al. 2011) also have several assumptions on their assessments including the expected known locations for tracking tower crane, or application of Scale Invariant Feature Transforms for initial recognition which limit their applications in actual construction operations. These approaches do not account for occlusions which is one of the main challenges on a dynamic construction site due to the dynamic nature of these environments. Most importantly 2D localization of the resources in most cases is still not fully automated. In a recent work, Park et al. (2011) proposed a method for 3D tracking the construction resources through a stereo camera, SIFT, and SURF (Speeded Up Robust Features)

detectors. Although this work mainly focuses on the 3D tracking based on correct recognition of resources in 2D, it does not propose any comprehensive approach for 2D learning and recognition of resources.

In the computer vision and robotics communities, there is a large number of emerging works in the area of tracking pedestrians using image sequences from fixed cameras, where background subtraction is mainly used to detect the objects (e.g., Zhao et al. 2008; Viola et al. 2003). More recent work has focused on applying trained pattern classifiers on individual video frames (e.g., Wu and Nevatia 2007; Seeman et al. 2007; Tuzel et al. 2007; Sabzmeydani and Mori 2007; and Dalal and Triggs 2005). The former group of research on pedestrian tracking is less related here as the background subtraction technique is used. In addition any temporal separation of video frames is not practical for the case of moving cameras (Bajracharya et al. 2009). These techniques particularly do not identify idle time which is a key information required for performance assessments. In case of the latter group, a variety of feature extraction and classification methods are used to achieve better performance (i.e., lower miss rates). Nonetheless, computational requirements are generally high or in several cases they are simply not benchmarked. Without a real-time detection of worker and equipment from video streams, detection of unsafe practices will not be effective.

2.3 Overview of the Proposed Method

Given 2D videos from fixed cameras, our goal is to automatically 1) detect and classify equipment in videos frames and sequences and 2) register their 3D locations. It is assumed that the video frames are expected to contain typical dynamic construction foregrounds and backgrounds that can generate occlusions.

Large variations in illumination, weather conditions, and resolution from one side in addition to the scale of equipment in 2D video streams and their intra-class variability makes site video streams challenging to work with. In order to tackle this problem, this research takes advantage from 1) sliding detection windows and 2) HOG (Dalal and

Triggs 2005) plus HOC descriptors to create our automated 2D detection method. For 3D localization the following steps are performed: 1) camera calibration (individual and stereo), 2) HOG+C feature matching, and 3) triangulation and non-linear optimization. These steps are described in the following subsections:

2.3.1 Equipment 2D Detection

1. Sliding Window Technique

The proposed method for detection of workers and equipment involves application of a detection sliding window. The basic idea is that the detection window scans across each video frame at all positions and spatial scales to find the best candidates for the resource. As shown in Figure 2.1 during this process, the detector window is tiled with a grid of overlapping blocks in which the features will be extracted. This strategy provides two key benefits: 1) detection of workers and equipment while idle, since it does not look into moving foreground objects, rather examines the possible candidates for their static representations and 2) detection of workers and equipment in close proximity of each other under high degrees of occlusions which is a key component for safety proximity assessments. In the following the process of detecting workers and equipment within each detector window is described.

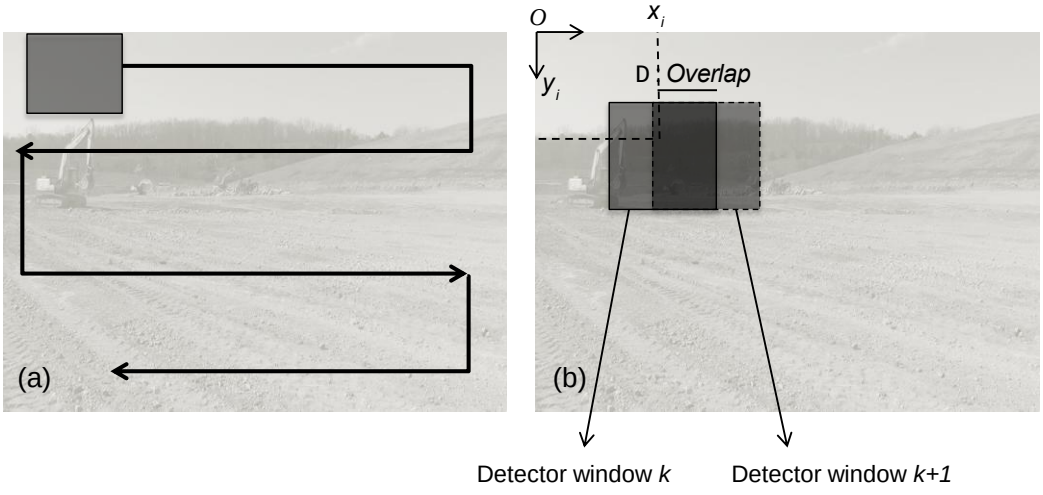


Figure 2.1: Representation of detection sliding window algorithm

2. Histogram of Oriented Gradients (HOG)

The main idea is that the local shape and appearance of equipment in a given detection window can be characterized by the distribution of local intensity gradients. The first step is to compute the magnitude $|\nabla f(x,y)|$ and orientation (angle) $\theta(x,y)$ of the gradient $\nabla f(x,y)$ at the normalized intensity of each detection window's pixel $I(x,y)$. Next, we derive the orientation Histogram of Gradients from these orientations and their magnitudes. The subimage (e.g., the shaded box in Figure 2.2a) covered by the detection window is divided into $t_x \times t_y$ overlapping blocks. Each block consists of l cells and each cell has $u \times v$ pixels (see Figure 2.2b). In each cell, the orientation histograms has n bins, which correspond to dominant of $t_x \times t_y$ orientations in the form of $i \times \pi / n$, $i = 1,...,n$ (see Figure 2.2c). The histogram computation involves distributing the weight of the orientation gradient magnitude for every pixel in the cell into the corresponding orientation bins. A naïve distribution scheme such as voting the nearest orientation bin would result in aliasing effect which is referred to distortion or artifacts due to under-sampling or poor reconstruction of digital video frames. Similarly, pixels near the cell boundaries would produce aliasing along spatial dimensions. Such aliasing effect can cause sudden changes in the computed feature vector. Similar to Burges (1998), the tri-linear interpolation of the pixel weight into the spatial orientation histogram is used to avoid this effect.

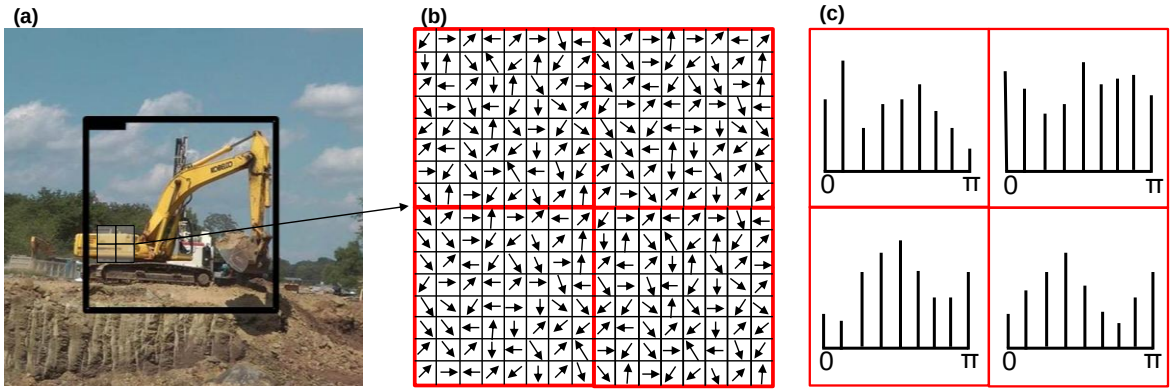


Figure 2.2: Histogram of oriented gradients: (a) a 250 x 250 detection window (the biggest square) in an image, (b) a 16 x 16 block consisting of 4 cells, and (c) the histogram of oriented gradients corresponding to the 4 cells.

3. Histogram of Colors (HOC)

Simultaneous to formation of HOG descriptors, the histograms of HS colors of the video frames are generated and concatenated with the HOG descriptors. As such, first the HS features are measured for the overall detection window. Next, the clustering technique is implemented on the HS features to cluster them into pre-determined cluster centers. We normalized the HS features to make them independent of different variation ranges and scales. Finally, the HOC are collected over the detector window and concatenated with the HOG. The HS descriptors are robust for construction scene saturated colors. Similar to (Weijer and Schmid 2006), it is hypothesized that adding HS colors in comparison can significant improve the detection and tracking of construction workers and equipment. This hypothesis is validated in (Memarzadeh et al 2012).

4. Support Vector Machine (SVM) Classifier

The last step is the machine learning process. For this purpose, we use a multiple and independent one-against-all Support Vector Machine (SVM) classification approach which each SVM is one of the margin-based classifiers [44]. Given n labeled training data $\{x_i, y_i\}$,wherein xi ($i = 1,...,n$, $x_i \in R^d$) is the probability distribution of the oriented gradients and colors for each video frame *i* with *d* dimensions, and $y_i \in \{0,1\}$ is the binary action class label (e.g., equipment or not-equipment), 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 is no prior knowledge about the distribution of the resource class video frames. Hence the optimal hyper plane is the one which maximizes the geometric margin γ as follows:

$$\gamma = \frac{2}{\|w\|} \tag{2.1}$$

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

$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$\text{subject to: } y_i\left(w.x_i + b\right) \geq 1 \text{ for } i = 1,...,N \tag{2.2}$$

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:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{subject to: } y_i\left(w.x_i + b\right) \geq 1 - \xi_i \text{ for } i = 1,...,N \tag{2.3}$$
$$\xi_i \geq 0 \text{ for } i = 1,...,N$$

In this formula, *C* represents a penalty constant which is determined by a cross-validation technique. In order to test the model and detect the resource classes, the classifier is extended into the form of several individual one-against-all classification schemes. Once the model is learned, the testing video frame datasets and the model are placed in the detection algorithm. The final outcome of this algorithm will be the resource classification results.

2.3.2 3D Localization of Resources

a. Camera Calibration

In order to perform a 3D localization of the detected construction equipment, after a comprehensive data collection process three main steps are taken: (1) camera calibration, (2) synchronizing detection windows across any given pair of video cameras, and (3) triangulation and 3D localization.

In order to know the true parameters of the cameras such as the position of the image center, focal length, scaling factors for pixels in cameras, skew factor, and lens distortion of each camera, cameras are at first calibrated using Camera calibration toolbox (Bouguet 2011). Camera calibration is a necessary step in 3D computer vision as it is used for 3D localization from two or more camera viewpoints. From a calibrated camera we can determine the distance of an object with respect to the location of the camera. Figure 2.3 shows the field engineer who's holding the camera calibration rig which will be conducted before and after every videotaping session. While performing data collection it is important to:

(a) keep the camera static (no changes in zoom, focus, or the location/viewing direction) during the data collection process,

(b) the camera positions should cover a wide field of view to enable tracking and localization of multiple operation equipment (see Figure 2.4), and

(c) there should be high percentage of overlap between video streams from multiple cameras. This strategy enables detection of salient feature points from various possible camera locations and viewpoint, and enables wide 3D baseline which further forms Epipolar geometry and support accurate tracking and localization of equipment (see Figure 2.4).

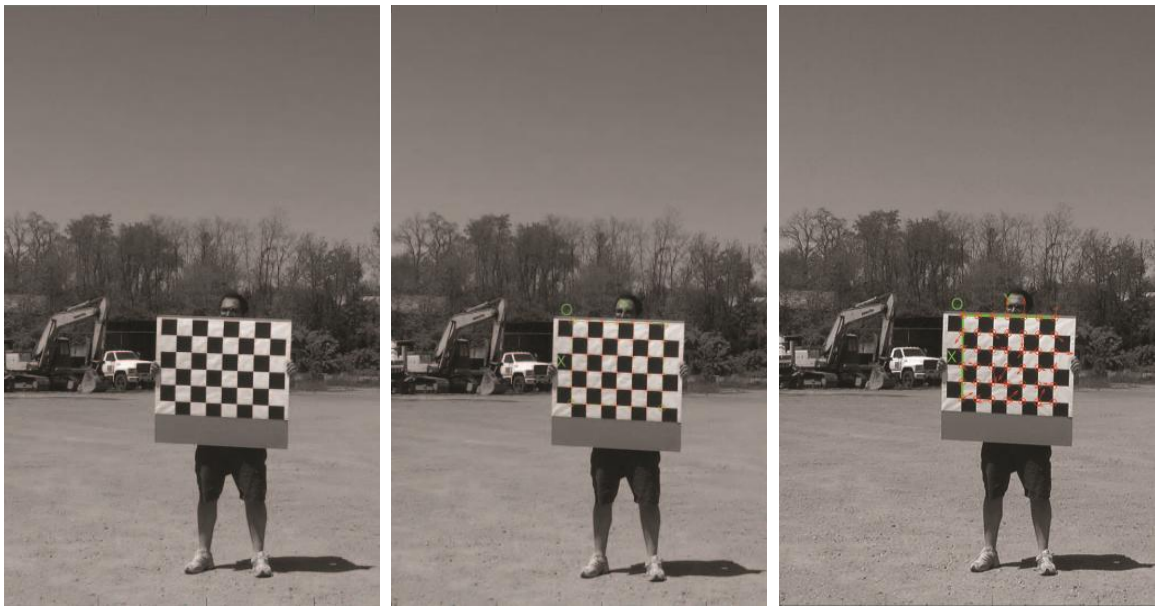


Figure 2.3: Field engineer performing camera calibration by moving the calibration rig around the frame in order to capture the most number of pixels for higher accuracy of 3D localization

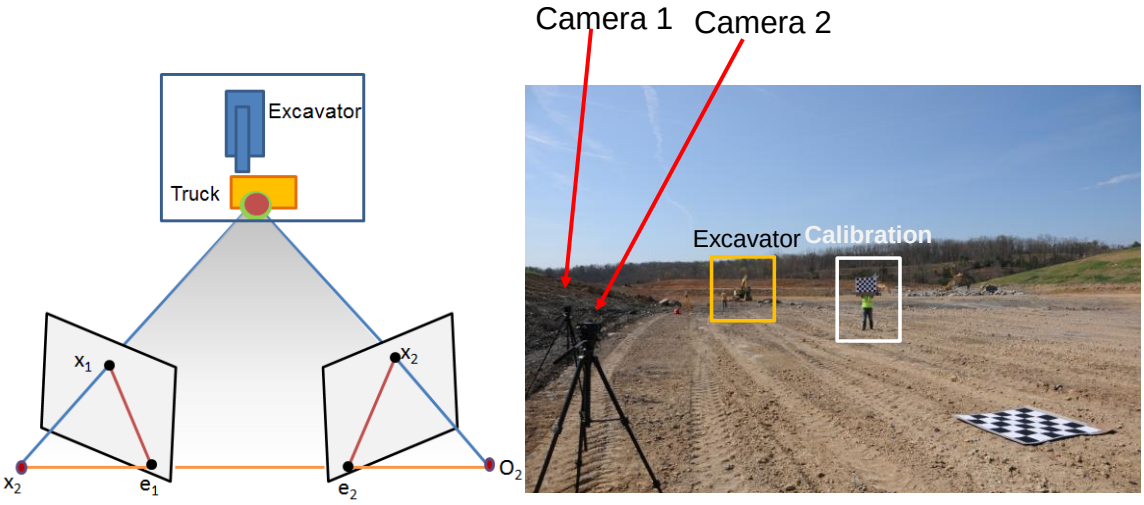


Figure 2.4: Epipolar Geometry

b. Individual Camera Calibration

To be able to conduct a projective mapping from 'world' coordinates (3D point position) to 'pixel' coordinates (2D point position), an individual camera calibration is performed. At first, the actual intrinsic and extrinsic parameters of each camera are measured

separately. Equation (3) shows the camera matrix where 2D point positions are used to represent the 3D point position:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} R & T \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{2.4}$$

where $\begin{bmatrix} u & v & 1 \end{bmatrix}^T$ is the 2D point position, $\begin{bmatrix} x_w & y_w & z_w & 1 \end{bmatrix}$ is the 3D point position, K is the intrinsic parameters; the extrinsic parameters R and T represent the rotation and translation of the camera respectively.

The intrinsic parameters define the internal camera parameters including (a) focal length, (b) principle point, (c) skew coefficient, and (d) distortions.

$$K = \begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.5}$$

where f_x and f_y represent the focal length in terms of pixels, γ is the skew coefficient, u_0 and v_0 are the principal points, which ideally are located in the center of the image. Figure 2.5 shows the estimated intrinsic re-projection error for one of the calibrated cameras, in which can be used to estimate the accuracy of the calibration process.

Figure 2.5: Camera calibration re-projection error

The extrinsic parameters R and T represent the coordinate system transformations from 3D coordinate system to 3D camera coordinate. Stereo calibration is used for this process which is explained in the next section.

c. *Stereo Camera Calibration*

In order to triangulate the location of the detected equipment in 3D, at first both cameras need to be brought up to the same coordinate system. Hence, for the case of a stereo configuration, the rotation and translation between the two cameras in their Epipolar geometry needs to be calculated. For this purpose, the stereo calibration toolbox of (Bouguet 2011) is used. Figure 2.6 shows the extrinsic parameters as a result of stereo calibration.

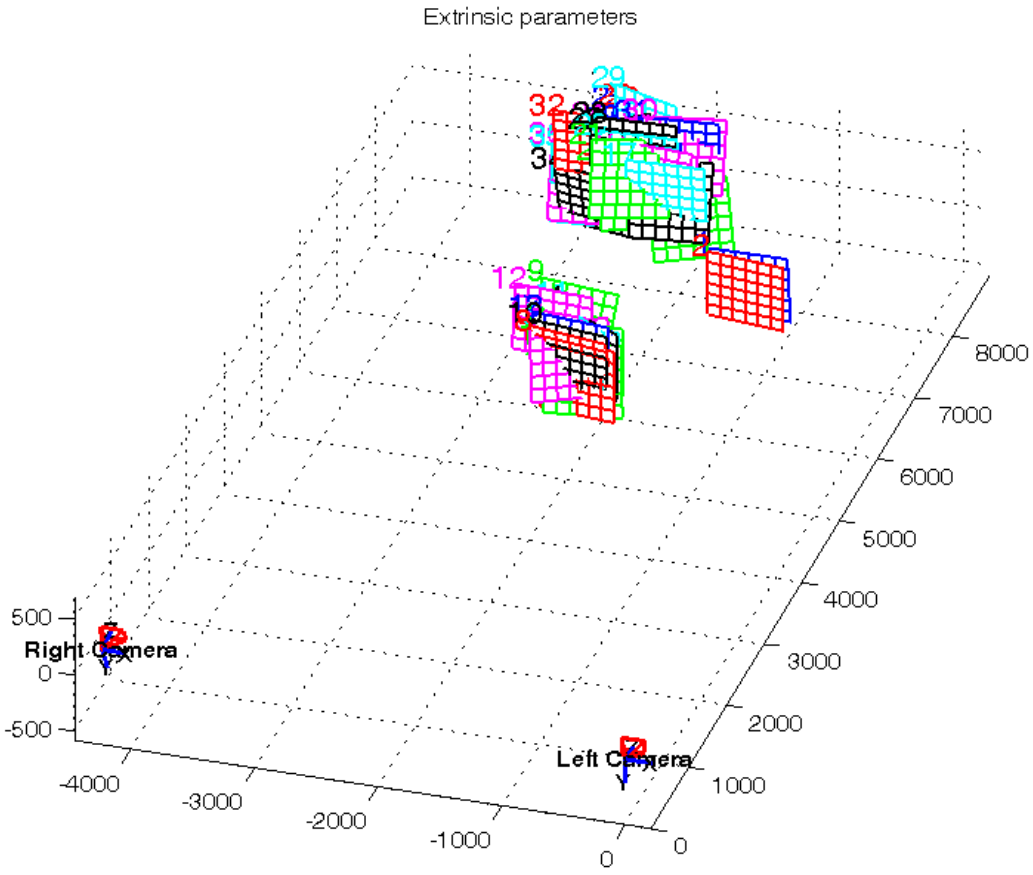


Figure 2.6: Extrinsic parameters calculated from the left and right cameras

d. Matching 2D Resources in Multiple Cameras

Once the HOG+C features are extracted for the 2D equipment detection, they are used for the matching between different frames. This process allows tracking equipment in 2D from a frame to the consecutive frames. To match the HOG+C feature descriptors, the minimum Euclidean distance between each pair of features is calculated. The process for the matching is similar to Lowe (2004), where the matching ratio test is conducted; i.e., for a feature descriptor in frame *i*, we find the two nearest neighbors in *j*, with distances d_1 and d_2, then accept the match if $d_1 / d_2 < 0.6$.

e. Triangulation and Non-Linear Optimization

To determine the 3D position of points in an image, a method known as triangulation (Hartley and Zisserman 2004) is used, given its position in two frames taken with cameras with known calibration and pose in 3D. Linear triangulation method is the direct analogue of the Direct Linear Transformation (DLT). In each image we have a measurement $x = MX$ for camera 1 and $x' = M'X$ for camera 2; these equations can be combined in the form of $AX = 0$, which is an equation linear in X.

First the homogeneous scale factor is eliminated by a cross product to give three equations for each image point, of which tow are linearly independent. For instance, for image 1, $x \times (PX) = 0$ and writing this out gives

$$\begin{aligned} x\left(p^{3T}X\right) - \left(p^{1T}X\right) &= 0 \\ y\left(p^{3T}X\right) - \left(p^{2T}X\right) &= 0 \\ x\left(p^{2T}X\right) - y\left(p^{1T}X\right) &= 0 \end{aligned} \tag{2.6}$$

where p^{iT} are the rows of *P*. These equations are linear in the components of *X*.

An equation of the form $AX = 0$ can then be composed, with:

$$A = \begin{bmatrix} xp^{3T} - p^{1T} \\ yp^{3T} - p^{2T} \\ x'p'^{3T} - p'^{1T} \\ y'p'^{3T} - p'^{2T} \end{bmatrix} \tag{2.7}$$

where two equations have been included from each image, giving a total of four equations in four homogenous unknowns. This is a redundant set of equations, since solution is determined only up to scale.

Singular value decomposition (SVD) can be looked at from three mutually compatible points of view. SVD is used to solve for the set of linear equations. The initial linear results are fed into non-linear Levenberg-Marquardt optimization.

2.4 Experimental Results and Validation

2.4.1 Data Collection and Experimental Setup

The data collection process of this approach consists of videotaping the construction equipment and comparing its actual location with those determined through the proposed algorithm. In this case, several pre-determined path for which the locations are properly surveyed (through GPS units) are determined and the operators of these equipment would be asked to move along these specified paths (Figure 2.7). During the experiments, the motion is compared with the trajectory identified through the algorithm. This validation is repeated for various cases to test the robustness of the approach for changes in scale, viewpoint, and also degrees of occlusion. Figure 2.8 shows a sample collected video displaying the path that the excavator is traveling.

Figure 2.7: GPS unit used to survey the points on the selected paths to benchmark the 3D localization results

Figure 2.8: Sample video frames demonstrating the excavator's path

2.4.2 Performance Evaluation Measures

To quantify and benchmark the performance of the action recognition algorithm, we plot the Precision-Recall curves and study the Confusion Matrix. These metrics are extensively used in the Computer Vision and Information Retrieval communities as set-based measures; i.e., they evaluate the quality of an unordered set of data entries. In the context of equipment action recognition, we define each as follows:

a. Precision-Recall Curve

To facilitate comparing the overall average performance of the variations of the proposed resource 2D tracking algorithm over a particular set of image datasets, 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 action instances that are

relevant to the particular classification, while recall is the fraction of relevant action instances that are retrieved. Thus, precision and recall are calculated as follows:

$$precision = \frac{TP}{TP + FP} \tag{2.8}$$

$$recall = \frac{TP}{TP + FN} \tag{2.9}$$

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 equipment detection window correctly recognizes equipment, it will be a TP; if a not-equipment instance is incorrectly recognized under equipment class, it will be a FP. When equipment instance is not recognized under the equipment 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 in form of a curve.

b. Confusion Matrix

DET curves present the same information as precision-recall graphs, yet allow small probabilities to be detected more easily. For this reason, majority of the human or pedestrian detection algorithms from the computer vision community are benchmarked and validated using these curves. Based on these DET 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:

$$miss\ rate = 1 - recall\ rate = \frac{\sum FN}{\sum (TP + FN)} \tag{2.10}$$

$$FPPW = \frac{\sum FP}{\sum (TN + FP)} \tag{2.11}$$

In several cases, average accuracy of the resource detection is also calculated using the following formula:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.12}$$

2.4.3 Experimental Results

a. Automated 2D Tracking

Throughout this section, we refer experimental results from our default proposed resource detector algorithms. Figure 2.9 shows several varying changes in our excavator database that were used to train the algorithm to automatically detect excavators from different viewpoints. As observed, our database for 2D tracking includes video frames of multiple resources with depicts the construction resource at different scales. Also several variations in pose, illumination, occlusion condition, and changes in the background are shown.

Figure 2.9: Example frames from video sequences of excavator operations. From left to right in rows: digging, hauling, dumping, and swinging action classes which illustrate tremendous appearance changes because of variability in equipment.

We implemented the proposed algorithms in MATLAB with several components in C++ for faster processing time. The implemented system was tested on a Linux 64bit platform with 24 GB RAM memory and 3.2 GHz Core i7 CPU. In our proposed method,

the RGB color space of the video frames is used with no gamma correction and the detectors have the following properties:

- The size of the detection windows for excavators are set to 250×250
- Linear gradient [-1;0;1] voting into 9 orientation bins in 0˚-180 ˚ is used for all cases; i.e., visually symmetrical gradients produce the best performance for detection of construction resources;
- L2-normalized blocks with 4 cells containing 8×8 pixels were used for all detection of excavators and finally,
- Linear SVM classifiers with C=1 are used for the detection and classification of the resources.

For the detection (testing phase), the detector window goes through the video frames at multiple uniform scales of the sizes (e.g., 1, 2, and 3). This strategy allows resources with smaller scales or within lower quality site video streams to be detected as well. Moreover, this technique helps the algorithm to be invariant to scale due to proximity of the resources to the video camera. Figure 2.10 shows the detected excavator and tracked in sequential video frames. Figure 2.11 shows the precision-recall curves for both HOG+HOC and HOG detectors and compares their performances for the excavator. As it is observed, the new method based on HOG and HOC descriptors significantly improves the performance of detecting construction resources.

Figure 2.10: Detecting excavators in sequential video frames.

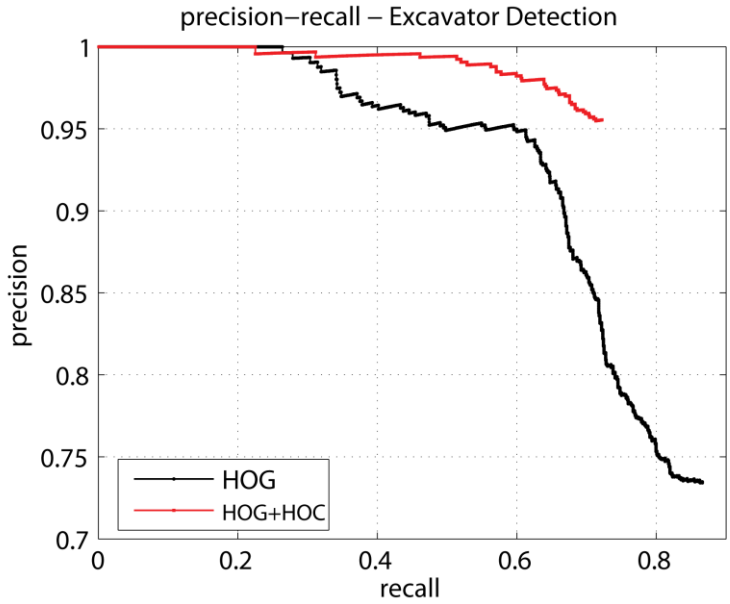


Figure 2.11: Overall results on performance of HOG and proposed HOG+C on detection of excavators

b. 3D Localization

After performing Direct Linear Transformation (DLT) followed by non-linear optimization, the Epipolar geometry is formed and the 3D location of the detected equipment is found. Figure 2.12a-f show the trajectory of the excavator movement. Figure 2.12g shows the overall trajectory of the excavator on the selected path. These coordinates are based on the center of each box in all the matched images. To validate this method, the trajectory results are compared with the GPS x,y, z coordinates with respect to the x,y,z coordinates of the cameras.

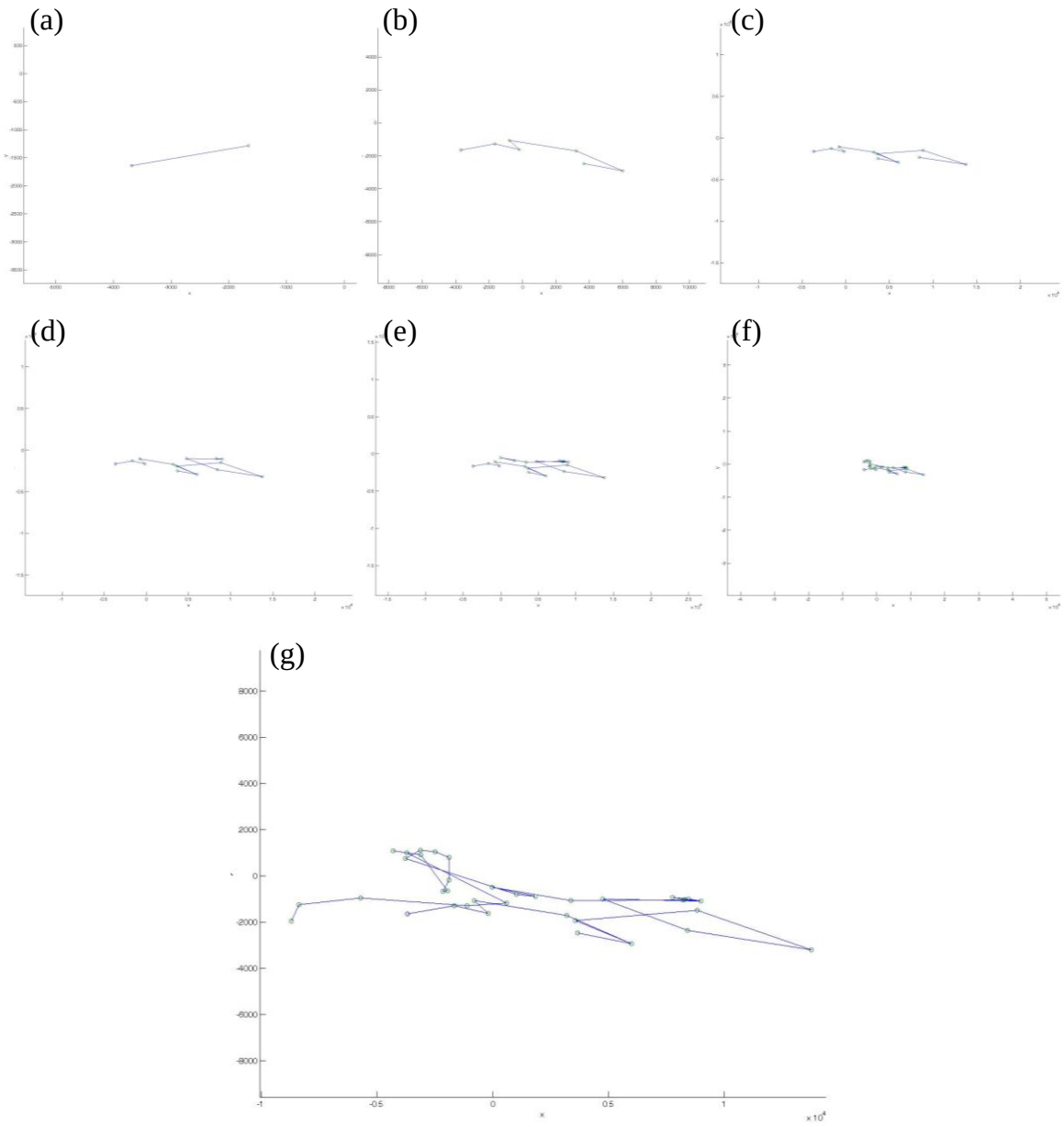


Figure 2.12: Excavator's movement trajectory

2.5 Discussion on the Proposed Method and Research Challenges

The average accuracies of the detection obtained for excavator is 98.83%. This performance is comparative to the state-of-the-art in both computer vision and AEC communities. In particular the ability to detect resources while idle shows superiority compared to previous methods presented in the AEC community. The presented results show the robustness of the proposed method to dynamic changes of illumination. The

proposed 3D localization algorithm is the first algorithm in the AEC community to use the 2D tracking results and through stereo calibration to determine the 3D position of the equipment in a given frame.

While this paper presented the initial steps towards processing site video streams for the purpose of 2D resource tracking and localization, several critical challenges remain. Some of the open research problems for our community include:

- **Real-time tracking in long video sequences.** Real-time and automated 2D tracking and localization of resources in long sequences of videos is a difficult task as like most sliding window algorithms, suffers from slow processing speed, making it unsuitable for safety proximity analysis. The 2D tracking and 3D localization algorithms presented in this paper are only capable of accurately tracking equipment in a post processing stage which limits their application for mainly performing action recognition To accurately track construction resources in real-time, more work is needed to implement the HOG+C based sliding window algorithm using the NVIDIA CUDA framework.

- **Variability in equipment types and models and worker body postures.** Accuracy of 2D detection is an important concern for applications such as productivity or safety proximity analysis. As such a comprehensive dataset of all types and models of equipment from all possible viewpoints is required for training purposes. The dataset presented in this work only includes two types of equipment from six different manufacturers. Development of larger datasets for equipment detection is still needed. In the case of construction workers, our dataset only included standing workers. Development of bending workers is also needed.

- **Temporal reasoning for 2D detection of resources.** Given the nature of construction project, it is very natural for construction resources to leave and come back to the field of view of a fixed camera on a jobsite. Also there might be

cases for which a resource is temporally fully occluded 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 tracking and localization using mobile cameras.** The ability to track construction workers and equipment from mobile cameras can open a lot of existing opportunities for context awareness of the resources on a jobsite. For example, a camera mounted on equipment can minimize the chances of accidents by eliminating the blind spots and alert the equipment 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 tracking resources using mobile cameras.

2.6 Conclusion

In this chapter, detail of 2D tracking and 3D localization of construction equipment is presented. In the proposed method, the 2D detection technique uses the histograms of oriented gradients and Hue-Saturation colors to initially detect the construction equipment in each camera and stores the HOG+C features. Through stereo camera calibration, the distance of the equipment with respect to the location of the cameras is measured. To determine the 3D position of the detected equipment in a video, DLT and non-linear optimization are used to form Epipoloar geometry.

2.7 Acknowledgements

The authors would like to thank the Virginia Tech Department of Planning, Design and Construction, Holder, and Skanska construction companies for providing access to their jobsites for a comprehensive data collection. The support of RAAMAC lab's current and former members, Chris Bowling and David Cline, Hooman Rouhi, Hesham Barazi, Daniel Vaca, Marty Johnson, Nour Dabboussi, and Moshe Zelkowicz is also appreciated.

The work is supported by a grant from Institute of Critical Technologies and Applied Science at Virginia Tech.

2.8 References

Bajracharya, M., Moghaddam, B., Howard, A., Brennan, S., and Matthies, L. H.(2009). "A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle". IJRR.

Bay, H., Ess, A., Tuytelaars, T., and Gool, L. V. (2008). "Speeded-Up Robust Features (SURF)." *Comput. Vis. Image Underst.*, 110(3), 346-359.

Bouguet, J.Y. (2004). "Camera calibration toolbox for Matlab." *Intel Corp.*, <http://www.vision.caltech.edu/bouguetj/calib_doc> (April 2011).

Brilakis, I., Park, M., and Jog, G. (2011). "Automated vision tracking of project related entities." *Advanced Engineering Informatics*, 25(4), 713-724.

Burges, C. J. C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Min. Knowl. Discov.*, 2(2), 121-167.

Caldas, C. H., Torrent, D. G., and Haas, C. T. (2006). "Using Global Positioning System to Improve Materials-Locating Processes on Industrial Projects." *Journal of Construction Engineering and Management*, 132(7), 741-749.

Cheng, T., Venugopal, M., Teizer, J., and Vela, P. A. (2011). "Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments." *Automation in Construction*, 20(8), 1173-1184.

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

Dalal, N., and Triggs, B. "Histograms of oriented gradients for human detection." *Proc., Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 886-893 vol. 881.

El-Omari, S., and Moselhi, O. (2009). "Data acquisition from construction sites for tracking purposes." *Engineering, Construction and Architectural Management*, 16(5), 490 - 503.

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

Fontana, R. J., Richley, E., and Barney, J. "Commercialization of an ultra wideband precision asset location system." *Proc., Ultra Wideband Systems and Technologies, 2003 IEEE Conference on*, 369-373.

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

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

Gong, J., and Caldas, C. H. (2008). "Data processing for real-time construction site spatial modeling." *Automation in Construction*, 17(5), 526-535.

Gong, J., and Caldas, C. H. "An Intelligent Video Computing Method for Automated Productivity Analysis of Cyclic Construction Operations." ASCE, 7-7.

Gong, J., and Caldas, C. H. (2010). "Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations." *Journal of Computing in Civil Engineering*, 24(3), 252-263.

Gong, J., Caldas, C. H., and Gordon, C. (2011). "Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models." *Advanced Engineering Informatics*, 25(4), 771-782.

Goodrum, P. M., Haas, C. T., Caldas, C., Zhai, D., Yeiser, J., and Homm, D. (2011). "Model to Predict the Impact of a Technology on Construction Productivity." *Journal of Construction Engineering and Management*, 137(9), 678-688.

Grau, D., Caldas, C. H., Haas, C. T., Goodrum, P. M., and Gong, J. (2009). "Assessing the impact of materials tracking technologies on construction craft productivity." *Automation in Construction*, 18(7), 903-911.

Hartley, R., and Zisserman, A. (2004). "Multiple view geometry in computer vision." *Cambridge University Press*.

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

Lowe, D. G. (2004). "Distinctive Image Features from Scale-Invariant Keypoints." *Int. J. Comput. Vision*, 60(2), 91-110.

National Institute of Science and Technology (NIST) (2011). "2011-2012 Criteria for Performance Excellence."

http://www.nist.gov/baldrige/publications/upload/2011_2012_Business_Nonprofit_Criteria.pdf (accessed September 2011).

Navon, R., and Sacks, R. (2007). "Assessing research issues in Automated Project Performance Control (APPC)." *Automation in Construction*, 16(4), 474-484.

Negri, P., Clady, X., Hanif, S. M., and Prevost, L. (2008). "A cascade of boosted generative and discriminative classifiers for vehicle detection." *EURASIP J. Adv. Signal Process*, 2008, 1-12.

Oglesby, C. H., Parker, H. W., and Howell, G. A. (1989). "Productivity Improvement in Construction." *McGraw-Hill, New York, NY* 84-130.

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

Rezazadeh Azar, E., and McCabe, B. (2011). "Automated Visual Recognition of Dump Trucks in Construction Videos." *Journal of Computing in Civil Engineering*, In Press.

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

Roweis, S. "Levenberg-marquardt optimization. http://www.cs.nyu.edu/~roweis/notes/lm.pdf. (accessed April 2012).

Sabzmeydani, P. and Mori, G. (2007). "Detecting pedestrians by learning shapelet features." *CVPR, IEEE.*

Seeman, E., Fritz, M. and Schiele, B. (2007). "Towards robust pedestrian detection in crowded image sequences." *CVPR, IEEE.*

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

Song, J., Haas, C. T., and Caldas, C. H. (2006). "Tracking the Location of Materials on Construction Job Sites." *Journal of Construction Engineering and Management*, 132(9), 911-918.

Su, Y. Y., and Liu, L. Y. "Real-Time Construction Operation Tracking from Resource Positions." ASCE, 25.

Teizer, J., Lao, D., and Sofer, M. (2007). "Rapid Automated Monitoring Of Construction Site Activities Using Ultra-Wideband." *The 24th International Symposium on Automation and Robotics in Construction. ISARC 2007*, Published by I.A.A.R.C., p.23-28.

Tuzel, O., Porikli, F. and Meer, P. (2007). “Human detection via classification on Riemannian manifolds.” *CVPR, IEEE.*

Viola, P., and Jones, M. "Rapid object detection using a boosted cascade of simple features." *Proc., Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, I-511-I-518 vol.511.

Weijer, J. v. d., and Schmid, C. (2006). "Coloring local feature extraction." *Proceedings of the 9th European conference on Computer Vision - Volume Part II*, Springer-Verlag, Graz, Austria, 334-348.

Williams, C., Cho, Y. K., and Youn, J.-H. "Wireless Sensor-Driven Intelligent Navigation Method for Mobile Robot Applications in Construction." ASCE, 76-76.

Wu, B. and Nevatia, R. (2007). “Simultaneous object detection and segmentation by boosting local shape feature based classifier.*” CVPR, IEEE.*

Yang, J., Vela, P. A., Teizer, J., and Shi, Z. K. "Vision-Based Crane Tracking for Understanding Construction Activity." ASCE, 32-32.

Zhai, D., Goodrum, P. M., Haas, C. T., and Caldas, C. H. (2009). "Relationship between Automation and Integration of Construction Information Systems and Labor Productivity." *Journal of Construction Engineering and Management*, 135(8), 746-753.

Zhao, Y., Gong, H., Lin, L., Jia, Y. (2008). “Spatio-temporal patches for night background modeling by subspace learning.” *ICPR*, Tampa, USA, 1-4.

Zou, J., and Kim, H. (2007). "Using Hue, Saturation, and Value Color Space for Hydraulic Excavator Idle Time Analysis." *Journal of Computing in Civil Engineering*, 21(4), 238-246.

Chapter 3: Automated Action Recognition of Earthmoving Equipment Using Vision-based Spatio-Temporal Features and Support Vector Machine Classifiers

3.1 Introduction

Equipment activity analysis, the continuous process of benchmarking, monitoring, and improving the amount of time construction equipment spend on different construction activities, can play an important role in improving construction productivity. This analysis examines the proportion of time equipment spend on different construction activities in a construction operation. A combination of detailed assessments and continuous improvements can help minimize the idle time, improve operational efficiency (Gong and Caldas 2010; Gong et al. 2011; Goodrum et al. 2011; Su and Liu 2007; Zhai et al. 2009), save time and money (Zou and Kim 2007), and result in a reduction of fuel use and emissions of construction operations (EPA 2010; Lewis et al. 2011). Through systematic implementation and reassessment, activity analysis can also extend equipment engine life and provide safer environments for equipment operators and workers.

Despite the benefits of activity analysis in identifying areas for improvement, an accurate and detailed assessment of work in-progress requires an observer to record and analyze the entire equipment's actions for every construction operation. Such manual tasks can be time-consuming, prohibitively expensive and prone to errors. In addition, due to the intra-class variability on how construction tasks are typically carried out, or in the duration of each work step, it is often necessary to record several cycles of operations to develop a comprehensive analysis of operational efficiency. Not only the traditional time-studies are labor intensive, but they also require a significant amount of time to be spent on manually collecting and analyzing data and can also affect the quality of the process as a result of the physical limitations or biases of the observer. Without a detailed and continuous activity analysis, it is unfeasible to investigate the relationship between the activity duty cycles versus productivity, or fuel use and emissions (Frey et al. 2010).

There is a need for a low-cost, reliable, and automated method for activity analysis that can be widely applied across all construction projects. This method needs to *remotely* and *continuously* analyze equipment's actions and provide detailed field data on their performance.

Over the past few years, cheap and high-resolution video cameras, extensive data storage capacities, and the availability of Internet connection on construction sites have enabled capturing and streaming construction videos on a truly massive scale. Detailed and dependent video streams provide a transformative potential for gradually and inexpensively sensing action and location of construction equipment, enabling construction companies to remotely analyze operational details and in turn assess productivity, emissions, and safety of their operations (Heydarian and Golparvar-Fard 2011). To date, the application of existing site video streams for automated performance assessment is still untapped and unexploited by researchers in most parts. A major reason is that these video streams are in forms that are not amenable for automated processing by traditional computer vision methods: the videos capture site operations from different camera locations and viewpoints and have wide variability and uncontrolled illuminations, resolution, and quality. The equipment type also has intra-class variability and the static and dynamic occlusions can significantly challenge development of automated computer vision based methods (static occlusion: the construction progress; dynamic: movement of other equipment and workers in a camera's field of view). One key challenge is *automated action recognition*; i.e., figuring out various actions equipment performs over time. While in the past year a few studies have looked into these areas (section 2), many challenging problems still remain unsolved.

This chapter focuses on the problem of automated action recognition for earthmoving equipment and a number of applications it enables. Figure 3.1 shows examples of the actions of an excavator and a truck operation, wherein the excavator performs a cycle of digging, hauling (swinging with full bucket), dumping, and swinging (with empty bucket) and the truck performs a cycle of filling, moving, and dumping.

Figure 3.1: Example frames from video sequences in excavator and truck action video datasets: Excavators: (a) digging; (b) hauling (swinging bucket full); (c) dumping; and (d) swinging (bucket empty); Trucks: (e) filling; (f) moving; and (g) dumping.

Given fixed cameras with small lateral movements, cluttered background, and moving equipment, the task is to automatically and reliably identify, categorize, and localize such actions. This paper presents an algorithm that aims to account for these scenarios. As such, the state-of-art research in this area is first overviewed. Next, a set of open research problems for the field, including action recognition under different camera viewpoints within dynamic construction sites are discussed. The new method expands on the work originally presented in (Heydarian et al. 2012) with significant algorithmic improvements on several parts and is accompanied with exhaustive validations. Also, a comprehensive dataset and a set of validation methods that can be used in the field for development and benchmarking of future algorithms are provided. The perceived benefits and limitations of the proposed method in the form of open research challenges are presented. Videos of the proposed method, along with additional supplementary material can be found at http://www.raamac.cee.vt.edu/equipmentactions.

3.2 Background and Related Work

In most state-of-the-art practices, the collection and analysis of the site performance data are not yet automated. The significant amount of information required to be manually collected may 1) adversely affect the quality of the analysis, resulting in subjective reports (Golparvar-Fard et al. 2009; Grau and Caldas 2009), and 2) minimize opportunities for continuous monitoring which is a necessary step for performance improvement (Golparvar-Fard et al. 2009; Gong and Caldas 2009; Grau and Caldas 2009;

Grau et al. 2009). Hence, many critical decisions may be made based on this inaccurate or incomplete information, ultimately leading to project delays and cost overruns.

In recent years, a number of research groups have focused on developing techniques to automatically assess construction performance. The main goal of these methods is to support improvement of operational efficiency and minimize idle times. Several studies such as (Gong and Caldas 2010; Gong et al. 2011; Goodrum et al. 2011; Su and Liu 2007) emphasize on the importance of a real-time resource tracking for improving construction performance. 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; Williams et al. 2007) , 3D range imaging cameras (Gong and Caldas 2008; Teizer et al. 2007), global and local positioning systems (GPS) (Gong and Caldas 2008; Teizer et al. 2007), and computer vision techniques (Brilakis et al. 2011; Park et al. 2011) have been tested to provide tracking data for onsite construction resources. While dominantly used for tracking construction material, they have also been used in locating workers and recording the sequence of their movement necessary to complete a task. Despite the benefits of location tracking for safety analysis, such methods do not provide enough information regarding operational performance of the equipment and workers. For performance assessment purposes, there is a need for automated recognition of resources' actions with reasonable accuracy.

3.2.1 Construction Equipment 2D and 3D Tracking

Several researchers including (Brilakis et al. 2011; Gong et al. 2011) have proposed vision-based methods for tracking project entities that have potential in addressing some of these requirements. However, current vision-based methods (e.g., (Gong et al. 2011; Zou and Kim 2007)) are either simulated in controlled environments or have only looked into automating one component of the overall method (mostly 3D tracking). Other vision-based location tracking approaches such as (Brilakis et al. 2011; Yang et al. 2011) have

several assumptions on their assessments including the expected known locations for tracking tower crane, or application of Scale Invariant Feature Transforms (SIFT) for initial recognition which limit their applications in actual construction operations. These approaches do not account for occlusions, which is one of the main challenges on construction sites due to the dynamic nature of these environments. In a recent work, (Park et al. 2011) proposed a method for 3D tracking the construction resources through a stereo camera, SIFT, and SURF (Speeded Up Robust Features) detectors. Their main focus is on 3D tracking based on correct recognition of resources in 2D and the work does not propose any comprehensive approach for recognizing and categorizing the visual appearance of the resources. Recent research proposes background subtraction on site video streams using several existing object recognition algorithms (Rezazadeh Azar and McCabe 2011) to differentiate between the moving and stationary objects (Chi and Caldas 2011). Background subtraction does not allow *idling* resources to be detected, which further limits their application for tracking and performance assessment purposes. Without a robust action recognition method, construction performance metrics cannot be measured. Any assumption that only uses location information to identify the type of action can be very misleading and does not provide enough information for the analysis of operational efficiency.

3.2.2 Construction Equipment Action Recognition

Despite a large number of emerging works in the area of human *action recognition* for smart online queries or robotic purposes and their significance for performance assessment on construction sites, this area has not yet been explicitly explored in the Architecture/Engineering/Construction (AEC) community. The work in (Gong and Caldas 2009) is one of the first in this area, which presented a vision-based tracking model for monitoring a tower crane bucket in concrete placement operations. The proposed method is mainly focused on action recognition of crane buckets and hence it cannot be directly applied to earthmoving operations. In a more recent work, (Gong et al. 2011) proposed an action recognition method based on an unsupervised learning method. The preliminary results are focused on three action categories of an excavator and four actions of workers (i.e., transporting, traveling, bending, nailing, and alignment). While

these representations indicate promising potentials, the actions are rather simple for performance assessments (e.g., actions such as concrete placement, concrete vibration, forming, finishing with hand tool, and finishing with machine tool need to be identified). In addition, due to the limited line of sight of a single camera, occlusions, varying illuminations, and in-class variability of workers and equipment (e.g., building construction excavators vs. mining excavators), the applicability of unsupervised learning models in unstructured construction sites can be challenging (This claim is validated in this paper). The work in (Zou and Kim 2007) also presented an image-processing approach that automatically quantifies the idle time of a hydraulic excavator. The approach uses color information for detecting motion of equipment in 2D and thus may not be robust to changes of scene brightness and camera viewpoint. The work is only focused on identifying non-idle/idle time and does not provide detailed information about various actions of the construction equipment which is necessary for performance assessment purposes.

Other major challenges in previous works on action recognition of workers and equipment include: 1) the lack of comprehensive video databases for action recognition of different types of equipment (considering different equipment types with various size, shape, and colors and videos taken from different distances and viewpoints); and 2) the use of controlled environments for training and testing of the proposed algorithms posing a problem toward handling more challenging situations such as multiple action recognition in dynamic construction environments.

3.2.3 Action Recognition in Computer Vision Community

In the computer vision community, there is a large number of researches in the area of person recognition and pose estimation (B. Yao and Fei-Fei 2011; Dalal and Triggs 2005; Dalal et al. 2006; Felzenszwalb et al. 2010; Wang et al. 2011; Yang and Ramanan 2011). The results of these algorithms seem to be both effective and accurate and in some cases (Felzenszwalb et al. 2010) they can also track deformable configurations which can be very effective for action recognition purposes. A number of approaches adopted visual representations based on spatio-temporal points (Dollar et al. 2005; Laptev 2005). This

can be combined with discriminative classifiers (e.g., SVMs) (Laptev et al. 2008; Marszalek et al. 2009), semi-latent topic models (Wang and Mori 2009), or unsupervised generative models (Niebles et al. 2008; Wong et al. 2007). Such approaches are effective but ignore temporal ordering of visual features in the video sequence. Other methods have shown the use of temporal structures for recognizing actions using Bayesian networks and Markov models (İkizler and Forsyth 2008; Laxton et al. 2007). To leverage the power of local features, (Niebles et al. 2008) introduced a new unsupervised model to learn and recognize the spatial-temporal features. Despite the benefits, this method requires labeled and segmented video sequences as an input. The work in (Savarese et al. 2008) introduced correlations that describe co-occurrences of code words within spatio-temporal neighborhoods. The size of the codebook, which is a set of representative spatio-temporal patterns, used in this work strongly influences the classification performance and the limited entries do not allow for good discrimination among these code classes. The work in (Liu and Shah 2008) determined the optimal size of the codebook using maximization of mutual information. This technique allowed two codebooks to be merged together if they have comparable distributions. To increase the precision of recognition with respect to space and time, (Yao and Zhu 2009) introduced an active basis of shape and flow patches. This technique requires minimal variance in the 2D locations of a resource within the video which can be limiting for construction applications.

Overall most existing computer vision methods require manual design and detailed trainings, which are time consuming. Also due to the nature of the experiments conducted, their application is mainly limited to simplified and controlled environments. Assumptions such as known starting points for each action within the videos and minimal acceptable variation in duration of each action can significantly impact the performance of these algorithms within less controlled video streams. While not readily available, certain elements of these works can be effectively used to create new methods suitable for equipment action recognition.

3.2.4 Limitations of Current Action Recognition Methods

Previous research on sensor-based approach has primarily focused on location tracking of workers and equipment, without paying much attention to monitoring their actions. In practice, when faced with the requirement for continuous benchmarking and monitoring of construction operations, there is a need for techniques that can support automated identification of construction actions. Site video streams offer great prospective for benchmarking and monitoring both location and action of construction resources. Despite the potential, similar to sensor-based approaches, most work on vision-based sensing in the AEC community has primarily focused on 3D location tracking of workers and equipment. The overall limitations of the state of the art computer vision approaches in action recognition are as follows:

1. Lack of systematic data collection and comprehensive datasets of action recognition of various construction equipment;

2. Lack of automated techniques that can detect articulated actions of construction equipment and workers plus their body posture necessary for performance assessments; (majority of vision-based approaches focus on recognizing simple actions; e.g., walking, jogging, running, boxing);

3. Assuming *a priori* for starting temporal point for each action in a temporal sequence. Without a proper knowledge on these starting points, a time-series of actions cannot be formed for further construction activity analysis;

4. None of the existing techniques look into simultaneous recognition of multiple actions, rather they look into simultaneous action recognition per single class of objects. For example, in pedestrian tracking, the focus is to detect a group action as opposed to multiple individual actions of pedestrians; and finally

5. None of the existing approaches takes a holistic approach to benchmarking, monitoring, and visualization of performance information. Without a proper

visualization, it will be difficult for practitioners to control the excessive impacts of performance deviations. In addition, understanding the severity levels of performance deviations will not be easy.

There is a need for techniques that can support automation of the entire process of benchmarking, monitoring, and control of performance deviations by identifying the sequence of resource actions, and determining idle/non-idle periods. Timely and accurate performance information brings awareness on project specific issues and empowers practitioners to take corrective actions, avoid delays, and minimize excessive impacts due to low operational efficiency (CII 2010). The proposed algorithm is presented in the following section.

3.3 Proposed Action Recognition Approach

Given a collection of site video streams collected from fixed cameras, our goal is to 1) automatically learn different classes of equipment actions present in the video dataset and 2) apply the leaned model to perform action recognition in new video sequences. The proposed approach is illustrated in Figure 3.2.

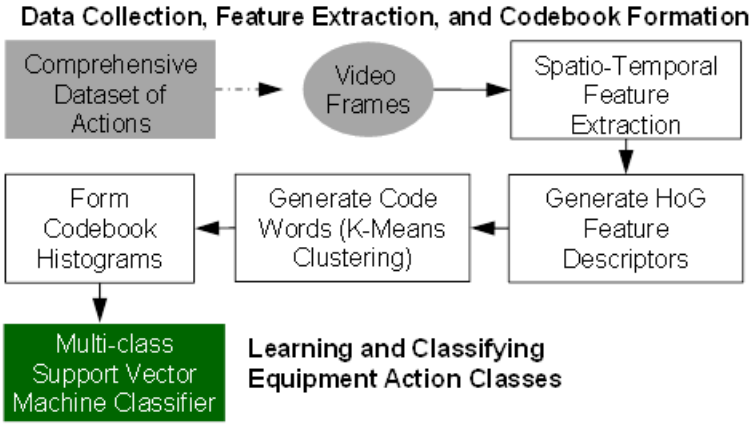


Figure 3.2: Flowchart of the proposed approach.

It is assumed that the videos can contain small camera motions such as those caused by the lateral movement of the camera due to wind. Also, the videos are expected to contain typical dynamic construction foregrounds and backgrounds that can generate motion clutter. In the training stage of our proposed method, it is assumed that each video

only contains one action of particular equipment. This assumption is relaxed at the full testing stage, where the proposed method can handle observations cluttered by the presence of other equipment performing various actions.

To represent all possible motion patterns for earthmoving equipment, a comprehensive video dataset for various actions is created. These videos, each containing single equipment performing only one action are initially labeled. First for each video, the local space-time regions are extracted using the spatio-temporal interest point detector (Dollar et al. 2005). A Histogram of Oriented Gradients (HOG) descriptors (Dalal and Triggs 2005) is then computed from each interest point. These local region descriptors are then clustered into a set of representative spatio-temporal patterns, each called code words. The set of these code words is from now on called a codebook. The distribution of these code words (is learned using a multi-class one-against-all Support Vector Machine (SVM) classifier. The learned model is then be used to recognize equipment action classes in novel video sequences. In the following each step is discussed in detail:

3.3.1 Feature Detection and Representation from Space-Time Interest Points

There are several choices in the selection of visual features to describe actions of equipment. In general, there are three popular types of visual features: static features based on edges and limb shapes (Feng and Perona 2002), dynamic features based on optical flow measurements (Dalal et al. 2006), and spatio-temporal features obtained from local video patches (Blank et al. 2005; Cheung et al. 2005; Dollar et al. 2005; Laptev 2005). Spatio-temporal features are shown to be useful in the articulated human action categorization (Niebles et al. 2008). Hence, in our method, videos are represented as collections of spatio-temporal features by extracting space-time interest points. To do so, it is assumed that during video recording, lateral movements do exist but are minimal. Our interest points are defined around the local maxima of a *response function.* To obtain the response, similar to (Dollar et al. 2005; Niebles et al. 2008) we apply 2D Gaussian and separable linear 1D Gabor filters as follows:

$$R = (I \otimes g \otimes h_{ev})^2 + (I \otimes g \otimes h_{od})^2 \tag{3.1}$$

where $I(x, y, t)$ is the intensity at location (x, y, t) of a video sequence, $g(x, y, \sigma)$ is the 2D Gaussian kernel, applied along the spatial dimensions, $h_{ev}(t; \tau, \omega)$ and $h_{od}(t; \tau, \omega)$ are the quadrature pairs of the 1D Gabor filter which are applied temporally.

$$g(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{x^2 + y^2}{2\sigma^2}\right\} \tag{3.2}$$

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega) \times \exp(-t^2 / \tau^2) \tag{3.3}$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega) \times \exp(-t^2 / \tau^2) \tag{3.4}$$

The two parameters σ and τ correspond to the spatial and temporal scales of the detectors respectively. Similar to (Dollar et al. 2005; Niebles et al. 2008), in all cases, $\omega = 4/\tau$ is used, and hence the response function R is limited to only two input parameters (i.e., σ and τ). In order to handle multiple scales of the equipment in the 2D video streams, the detector is applied across a set of spatial and temporal scales. For simplicity in the case of spatial scale changes, the detector is only applied using one scale and thus the codebook is used to encode all scale changes that are introduced and observed in the video dataset; i.e., our video dataset contains multiple spatial scales of each equipment for training purposes). It is noted in (Dollar et al. 2005; Niebles et al. 2008) that any 2D video region with an articulated action can induce a strong response to the function R. This is due to the spatially distinguishing characteristics of actions, and as a result those 2D regions that undergo pure translational motion or do not contain spatially distinguishing features will not induce strong responses. The space-time interest points are small video neighborhoods extracted around the local maxima of the response function. Each neighborhood is called a *cuboid* and contains the local 3D video volume that contributed to the response function (3rd dimension is time). The size of the cuboid is chosen to be six times the detection scales along each dimension ($6\sigma \times 6\sigma \times 6\tau$). To obtain a descriptor for each cuboid, a Histogram of Gradients (HOG) (Laptev et al. 2008) is then computed. The detailed process is as follows:

At first, the normalized intensity gradients on *x* and *y* directions are calculated and the cuboid is smoothed at different scales. Here the normalized intensity gradients are representing the normalized changes of the average intensities, and the smoothing is conducted using the response function *R*. The gradient orientations are then locally histogrammed to form a descriptor vector. The size of the descriptor is equal to (the number of pixels in the cuboid) × (the number of temporal bins) × (the number of gradients directions). In our case, this descriptor size is $(3\times3)\times2\times10=180$. In addition to the application of HOG descriptors, histograms of optical flow (Efros et al. 2003) was also considered. As validated in section 3.4, the HOG descriptor results in superior performance. Figure 3.3 shows an example of interest points detected for an excavator's 'digging' action class. Each small box represents a detected spatio-temporal interest point. Figure 3.4 shows an example of the HOG descriptor for one of the interest points from the excavator's digging action class.

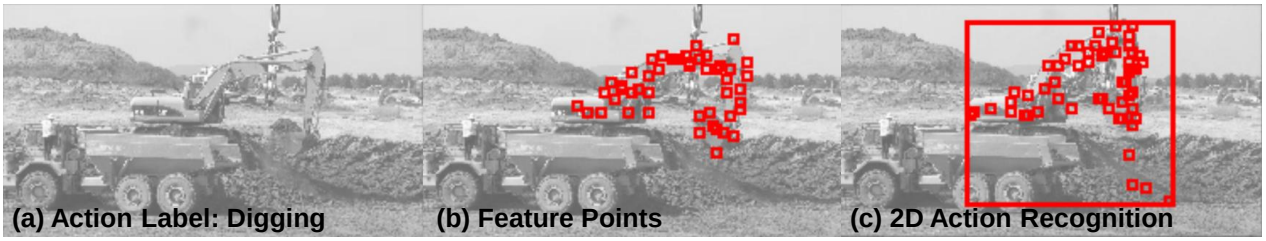
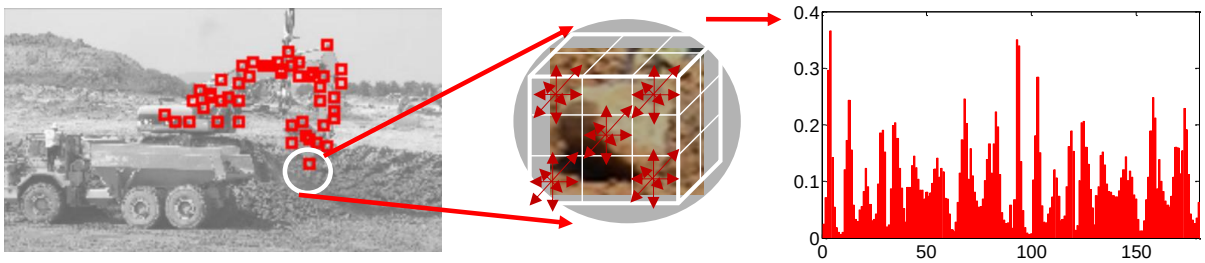


Figure 3.3: Detection of the spatio-temporal features. Each small box in (b) and (c) corresponds to a cuboid that is associated with a detected interest point. The 3-dimensions of each cuboid are size times scale parameters σ and τ of the detector. (c) shows the final outcome of the action recognition and localization (Figure best seen in color).



(a) Spatio-Temporal Features (b) Magnitude and Orientation Of Intensity Gradients (c) 180-bin HOG Descriptor

Figure 3.4: HOG descriptor for one spatio-temporal feature from one video of the excavator's digging action class dataset.

3.3.2 Action Codebook Formation

In order to learn the distribution of spatio-temporal features in a given video, first a set of HOG descriptors corresponding to all detected interest points in the entire training video dataset is generated. Using the *k*-means clustering algorithm and the Euclidean distance as the clustering metric, the descriptors of the entire training dataset are clustered into a set of code words. This result of this process is a codebook that associates a unique cluster membership with each detected interest point. Hence, each video is represented as a distribution of spatio-temporal interest points belonging to different code words. Figure 3.5 illustrates the action codebook formation process. A total of 350 cluster centers are considered for the best action recognition performance. The effect of the codebook size (the number of code words) and its impacts on the action classification accuracy are explored in section 4.3.3 of this paper.

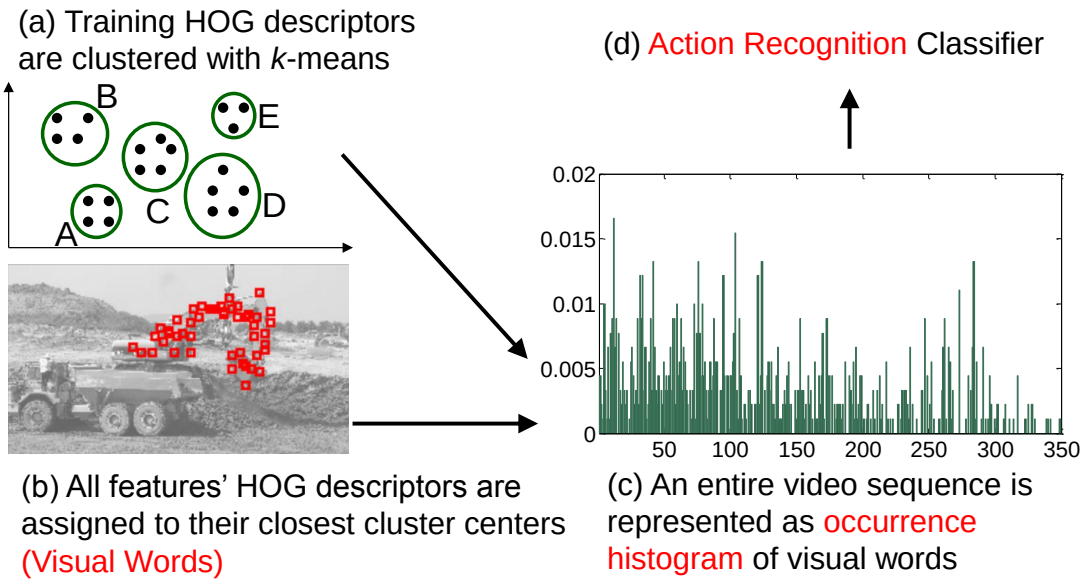


Figure 3.5: Action recognition codebook formation process.

3.3.3 Learning the Action Models: Multi-class One-Against-All Support Vector Machine Classifier

To train the learning model of the action categories, a multi-class one-against-all Support Vector Machine (SVM) classifier is introduced. The SVM is a discriminative machine learning algorithm which is based on the structural risk minimization induction principle (Vapnik and Bottou 1977). In this work, it was hypothesized that traditional classifiers such as Naïve Bayes (Rish 2001) or unsupervised learning methods such as probabilistic Latent Semantic Analysis (pLSA) (Hofmann 1999) may not obtain the best recognition performance. For equipment action classification, the number of samples per class can be limited and consequently these methods tend to result in over-fitting. In the following, these algorithms are briefly introduced. The performance of these algorithms for learning equipment action classes is compared in section 3.3.4.d and the hypothesis for application of a multiple one-against-all supervised SVM classifier is validated.

a. The Multiple Binary Support Vector Machine Classifier

In the proposed multiple one-against-all SVM classifier, for each action category (k), a separate binary-class linear kernel SVM (Ψ_k) is built so that video instances associated with that label are within the same class and the rest of the videos are in another. This

casts the problem into a one-against-all classification scheme. For example, one of the binary SVM classifiers decides whether a new excavator video belongs to the '*Digging*' or '*non-Digging*' action classes. Given N labeled training data $\{x_i, y_i\}, i = 1,...,N$, $y_i \hat{I} \{0,1\}$, $x_i \hat{I} R^d$, wherein x_i is the probability distribution of the spatio-temporal interest points for each video (i) with d dimensions (occurrence histograms of visual words), and y_i is the binary action class label, 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 there is no prior knowledge about the distribution of the action class videos. Hence the optimal hyper plane is the one which maximizes the geometric margin g (Burges 1998) as follows:

$$\gamma = \frac{2}{\|w\|} \tag{3.5}$$

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

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &\text{subject to: } y_i(w.x_i + b) \geq 1 \text{ for } i = 1,...,N \end{aligned} \tag{3.6}$$

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

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ &\text{subject to: } y_i(w.x_i + b) \geq 1 - \xi_i \text{ for } i = 1,...,N \\ &\qquad\qquad \xi_i \geq 0 \text{ for } i = 1,...,N \end{aligned} \tag{3.7}$$

In this formula, C represents a penalty constant which is determined by a cross-validation technique. In order to test the model and recognize the equipment action

classes, the classifier is extended into the form of a multi-class one-against-all classification scheme. For each binary classifier, the action classes in addition to the classification decision scores are stored. Among all binary classifiers, the one which results in the highest classification score is chosen as the equipment action class, and the outcome of each video's classification is labeled accordingly.

b. Naïve Bayes Classifier

A Bayes classifier is a simple probabilistic classifier based on applying Bayes' statistics theorem with strong (naive) independence assumptions. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the classifier's parameters. The probability model for a Naïve Bayes classifier $P\left(C\middle|W_1,...,W_n\right)$ over a dependent class variable *C* is a model with a small number of equipment action classes, conditional on several feature variables W_1 through *Wd*. Here the feature variables are the *d* code words of the occurrence histograms. Assuming that each feature variable W_i *is* independent; the conditional distribution over the class variable can be expressed as:

$$P\left(C\middle|W_1,...,W_n\right)=\frac{1}{\mathrm{K}}P(C)\prod_{1}^{n}P(W_i\,|\,C) \tag{3.8}$$

where K is a scaling factor dependent only on $W_1,...,W_n$; i.e., a constant if the values of the feature variables are known. In the case of equipment action classification, these features are considered to have similar impacts and hence K is '*d*'.

c. Probabilistic Latent Semantic Analysis (pLSA) Classifier

Probabilistic Latent Semantic Analysis (pLSA) is a statistical technique for the analysis of binary-mode and co-occurrence data, proposed in (Hofmann 1999) and in recent years has been extensively used for text, object, and human action recognitions. In text recognition, each document is generatively modeled as a bag of codewords (bag-of-words), each sampled from a document-specific mixture of *Z* latent 'topic' distribution. Each topic *z* is described by its distribution *p(w|z)* over the *W* possible words of the

dictionary and each document *d* is characterized by the mixture over *Z* topics. Figure 3.6 shows a graphical for the pLSA model.

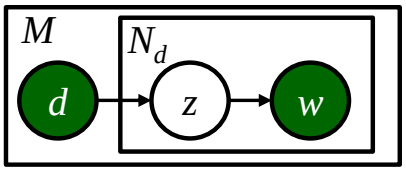


Figure 3.6: The probabilistic Latent Semantic Analysis (pLSA) model. This figure is reproduced from (Niebles et al. 2008)

Here *d* represents video sequences, *z* represents equipment action categories, and *w* represents the code books. The *n* code words of a video sequence *d* are treated as a set of independent parameters. Letting *z* denote the unknown action of code book W_b, the joint probability of the *n* code words and corresponding *d* is modeled as follows:

$$\begin{aligned} p(w_1, w_2, ..., w_n, d) &= \prod_{i=1}^{n} p(w_i \mid d)\, p(d) \\ &= \prod_{i=1}^{n} \sum_{z \in Z} p(w_i \mid z)\, p(z \mid d)\, p(d) \end{aligned} \tag{3.9}$$

In the case of equipment action recognition, the action video streams are considered as a mixture of topics and local patches often produced by some interest point detectors, and are viewed as visual words. Thus equipment action snapshots are modeled as a mixture of latent topics that generates each patch independently.

3.4 Experimental Results and Validation

3.4.1 Data Collection and Experimental Setup

Due to the lack of databases for training visual actions of different construction equipment, before testing our algorithm, it was necessary to create a comprehensive benchmarking video dataset. Given the variety of the form and shape for construction equipment and due to different representations from different camera viewpoints, different lighting and weather conditions, and finally static and dynamic occlusions, it is very important to assemble a comprehensive action recognition dataset. Since the focus of this paper is on earthmoving operations, a comprehensive dataset for several types of

equipment is formed. Upon successful evaluation, datasets for other construction equipment can be collected and the proposed method can be tested on those for action recognition purposes. Particularly the following combinations of equipment are considered in our database: 1) excavators and dump trucks, 2) backhoes and dump trucks, 3) scrapers, excavators, and dump trucks, 4) scrapers, dozers, and dump trucks, and finally 5) loaders and dump truck. For each of these combinations, we recorded a distinct video database containing all possible actions of the equipment. For example, for the combination of excavators and dump trucks, this video database contains five types of excavator actions (i.e., moving, digging, hauling [swing with full bucket], swinging [empty bucket], and dumping) and three types for dump truck actions (i.e., moving, filling, and dumping). This dataset contains three types of excavators (manufacturers: Caterpillar, Komatsu, and Kobelco) and three types of dump trucks (manufacturers: Caterpillar, Trex, and Volvo). 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. To ensure various types of backgrounds and level of occlusions, the videos were collected from five different construction projects (i.e., two building and three infrastructure projects). Due to various possible appearances of equipment, particularly, their actions from different views and scales in a video frame, as shown in Figure 3.7, several cameras were set up in two 180° semi-circles (each camera roughly 45° apart from one another) at the training stage. This strategy enables the equipment 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 codebooks, all possible scales are considered.

Overall a total of 150 to 170 training videos were annotated and temporally segmented for each action of equipment (overall 895 videos for four and three action classes of excavators and dump trucks). Each video has different durations, and hence various possible temporal scales for each action are introduced into the training dataset. The “idle” action category is not used for training purposes. Rather idle time frames are determined when there are no spatio- temporal features detected for a given number of

consecutive frames. The video dataset is made public at: (www.raamac.cee.vt.edu/equipmentactionrecognition).

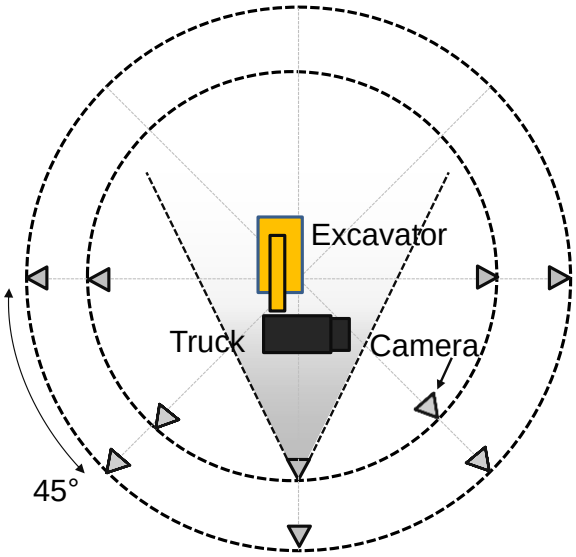


Figure 3.7: Data Collection and Experimental Setup.

3.4.2 Performance Evaluation Measures

To quantify and benchmark the performance of the action recognition algorithm, we plot the Precision-Recall curves and study the Confusion Matrix. These metrics are extensively used in the Computer Vision and Information Retrieval communities as set-based measures; i.e., they evaluate the quality of an unordered set of data entries. In the context of equipment action recognition, we define each as follows:

a. Precision-Recall Curve

To facilitate comparing the overall average performance of the variations of the proposed action recognition algorithm over a particular set of equipment action datasets, individual action 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 action instances that are relevant to the particular classification, while recall is the fraction of relevant action instances that are retrieved. Thus, precision and recall are calculated as follows:

$$precision = \frac{TP}{TP + FP} \tag{3.10}$$

$$recall = \frac{TP}{TP + FN} \tag{3.11}$$

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 a *digging* video is correctly recognized under the *digging* action class, it will be a TP; if a dumping video is incorrectly recognized as digging, it will be a FP for the digging class. When a *digging* video is not recognized under the digging action 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 action class for any recall level great than or equal to *i*. For each recall level, the precision is calculated, and then the values are connected and plotted in form of a curve.

b. Confusion Matrix

The performance of the action classifiers (i.e., *digging*, *dumping*, and *hauling* classifiers for excavator video dataset) is analyzed using the confusion matrix. The confusion matrix returns the average accuracy per action class. The average accuracy of the action classification is calculated using the following formula:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.12}$$

A confusion matrix shows for each pair of action classes $<c_1, c_2>$, how many action videos from c_1 were incorrectly assigned to c_2. Each column of the confusion matrices represents the *predicted* action class and each row represents the *actual* action class. The detected TPs and FPs are compared and the percentage of the correctly predicted classes with respect to the actual class is created and represented in each row.

3.4.3 Experimental Results

In this following section, we first present the experimental results from our proposed algorithm. Then, in the subsequent sections, we test the efficiency of our approach for the recognition task on various model parameters; i.e., feature detection, feature descriptors, codebook sizes, and finally the machine learning classifier.

For the excavator and dump truck actions datasets, which contain 626 and 233 short single-action sequences respectively, the interest points were extracted and the corresponding spatio-temporal features described using the procedure described in section 3.1. Some sample video frames from different equipment actions with scale, viewpoint, and background changes are shown in Figure 3.8.

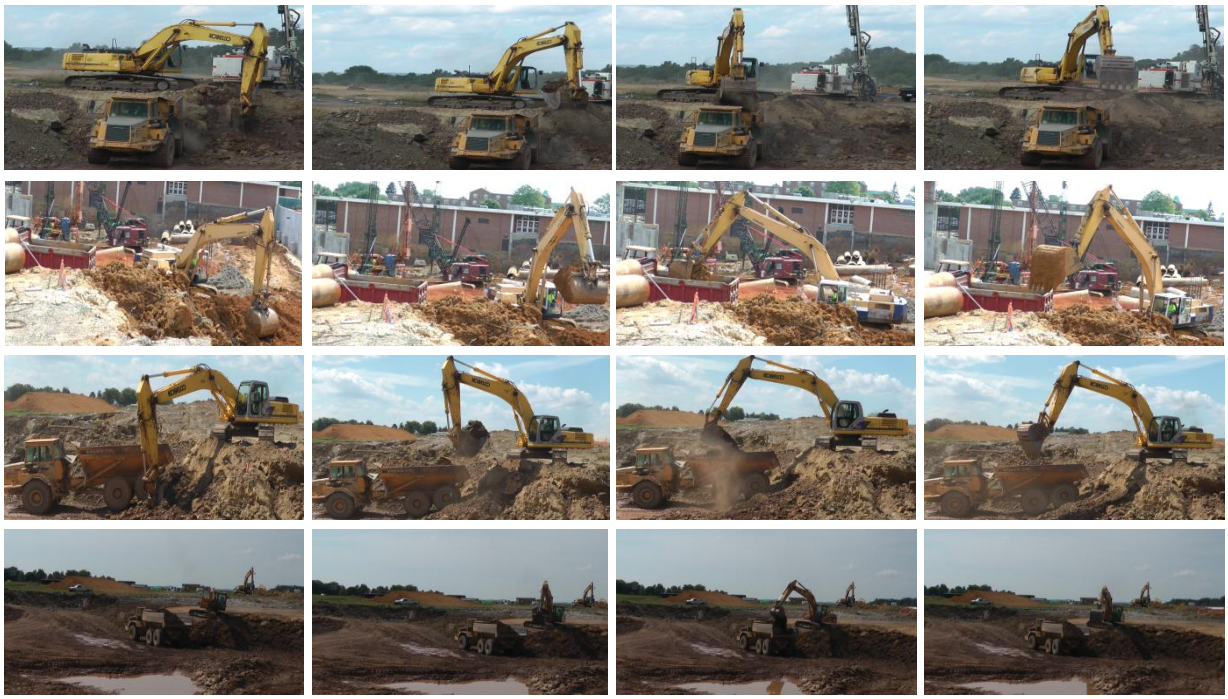


Figure 3.8: Snapshots from different actions of an excavator's operations. The dataset contains four types of actions. These actions are recorded from Caterpillar, Komatsu, and Kobelco models of excavators in different construction sites from various viewpoints and at different scales. The camera has minor lateral movement and in several cases, the foreground and background contains other movements.

The detector parameters are set to $\sigma = 1.5$ and $\tau = 3$ and histograms of gradients (HOG) are used to describe the feature points. Some examples of the detected spatio-

temporal feature patches are shown in Figure 3.9. Each row represents the number of video frames that are used to describe the feature. In order to form the codebook, 350 code words (*k*-means cluster centers) were selected and the spatio-temporal features for each video were assigned to different code words. The outcome is the codebook histogram for each video. Next, we learn and recognize the equipment action classes using the multi-class linear SVM classifiers. To solve Eq. 3.7, we use the libSVM (Chang and Lin 2011) and set the kernel type to *C-SVC*. For each action classifier, a decision value is learned. Comparing these decision values enables the most appropriate action class to be assigned to each video.

In order to test the efficiency of our approach for the action recognition task, we divided the action dataset into training and testing sequences with a ratio of 3 to 1 and computed the confusion matrix for evaluation. This process of splitting training and testing videos randomly is conducted five times, and the average precision values are reported for the confusion matrix. The algorithms were implemented in Linux 64bit Matlab on an Intel Core i7 workstation laptop with 8 GB of RAM.

Figure 3.9: Each row contains the frames from the neighborhood of a single spatio-temporal interest point which is assigned to different action categories.

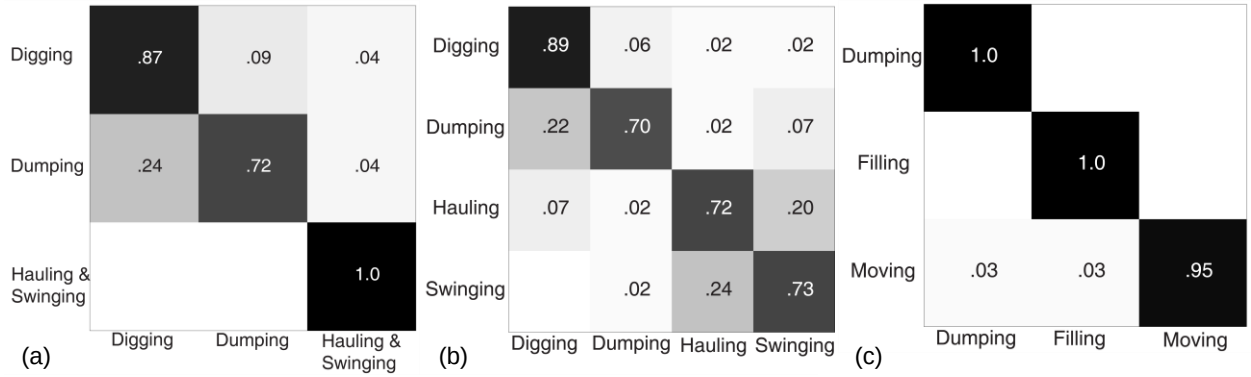


Figure 3.10: (a) and (b) Confusion matrix for excavator's three and four-action class datasets (average performance = 86.33% and 76.0% respectively; (c) Confusion matrix for dump truck dataset (performance average = 98.33%).

For excavator action recognition, Figure 3.10a shows that the largest confusion happens between 'hauling' and 'swinging' action classes. This is consistent with our intuition that both these actions are visually similar (hauling: bucket full vs. swinging: bucket empty). Hence we combined these actions classes assuming that in longer video sequences, the order of equipment actions can help easily distinguish them from one another; i.e., hauling can only happen after digging is detected. Figure 3.10b shows the recognition performance for excavators when three action classes are considered. Another significant confusion occurs between 'digging' and 'dumping' action classes. These actions are also visually similar (bucket getting closer or farther from the excavator arms). Figure 3.11 shows the decision values and the action classification results for the entire equipment action dataset. The horizontal axis in these figures represents the video dataset (see

Table 3.1 for details of each action class). The hyper plane in each individual binary classification s automatically learned through the binary linear SVM classifier. As observed in Figure 3.11.d, e, and f, the most appropriate action class can be selected for each video by comparing the decision values from the binary classification results.

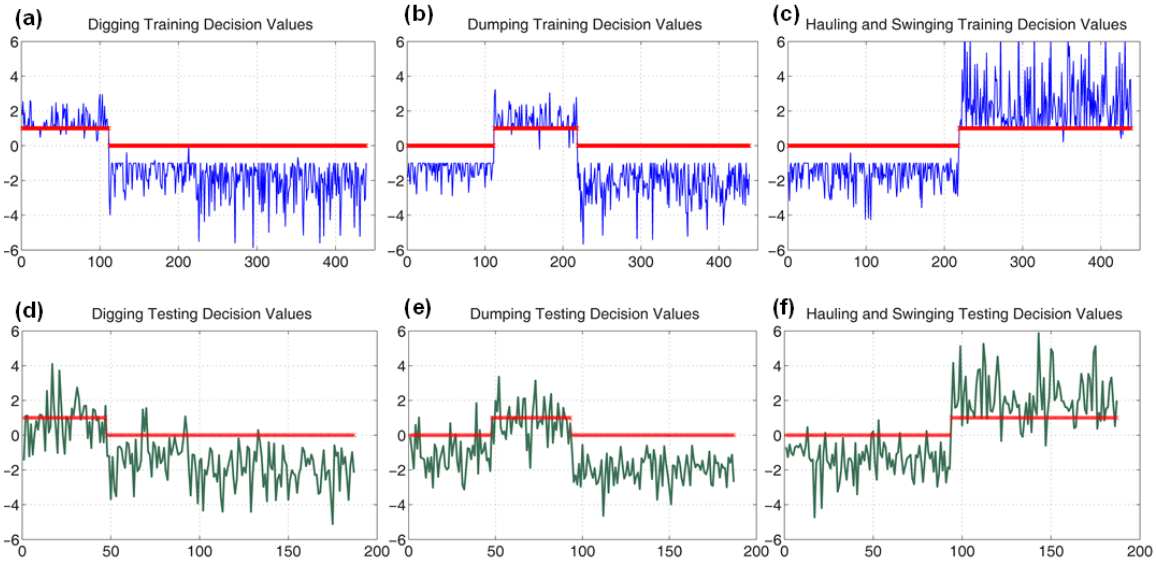


Figure 3.11: Decision Values for both training and testing of the linear SVM classifiers. Each row from left to right shows the values for 'Digging', 'Dumping' and combined 'Hauling and Swinging' decision values for all video instances.

Table 3.1: Excavator and truck action classification datasets.

Equipment	# of Videos	Action Class	Training	Testing
Excavator	159	Digging	#0-110 (111)	#0-47 (48)
	153	Dumping	#111-217 (107)	#48-94 (46)
	315	Hauling/Swinging	#218-437 (220)	#95-190 (95)
Truck	85	Filling	#0-58 (59)	#0-25 (26)
	126	Moving	#59-146 (88)	#26-64 (38)
	22	Dumping	#147-161 (15)	#65-71 (7)

Figure 3.12 shows the precision-recall curves for the action classification of the excavator and truck testing dataset using the multi-class binary SVM classifiers. The combined hauling and swinging action class for the excavator action recognition and the moving class for the dump truck action recognition have the best average performances.

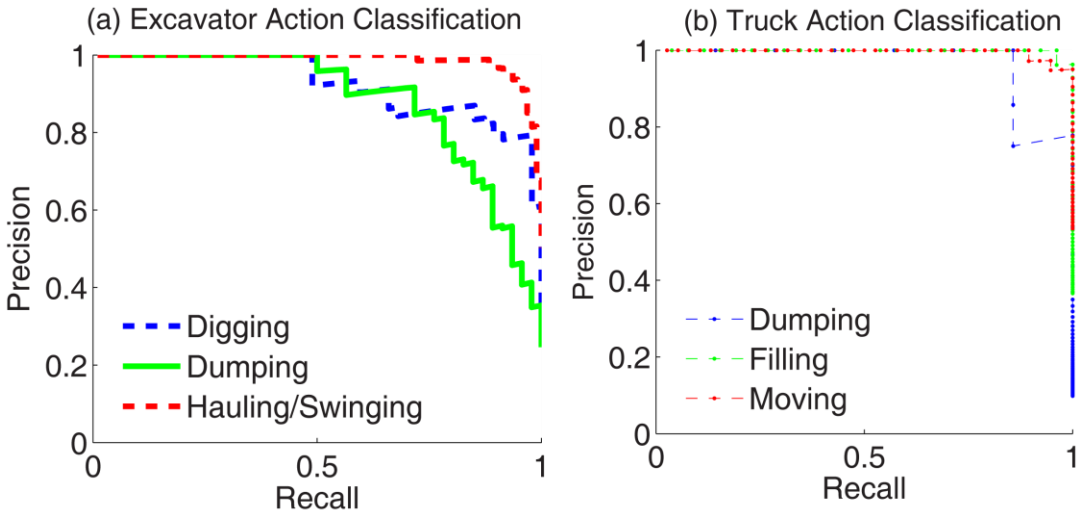


Figure 3.12: Precision-Recall curves for excavator and truck action classifications.

Several example features from the testing sequences in both truck and excavator video datasets are shown in Figure 3.13. In this figure, each spatio-temporal feature patch is automatically color coded with the corresponding action category. Also note that in several videos of truck and excavator datasets (e.g., Figure 3.13 a-4, b-2, and b-3), the spatio-temporal features from partially occluded equipment are detected and categorized.

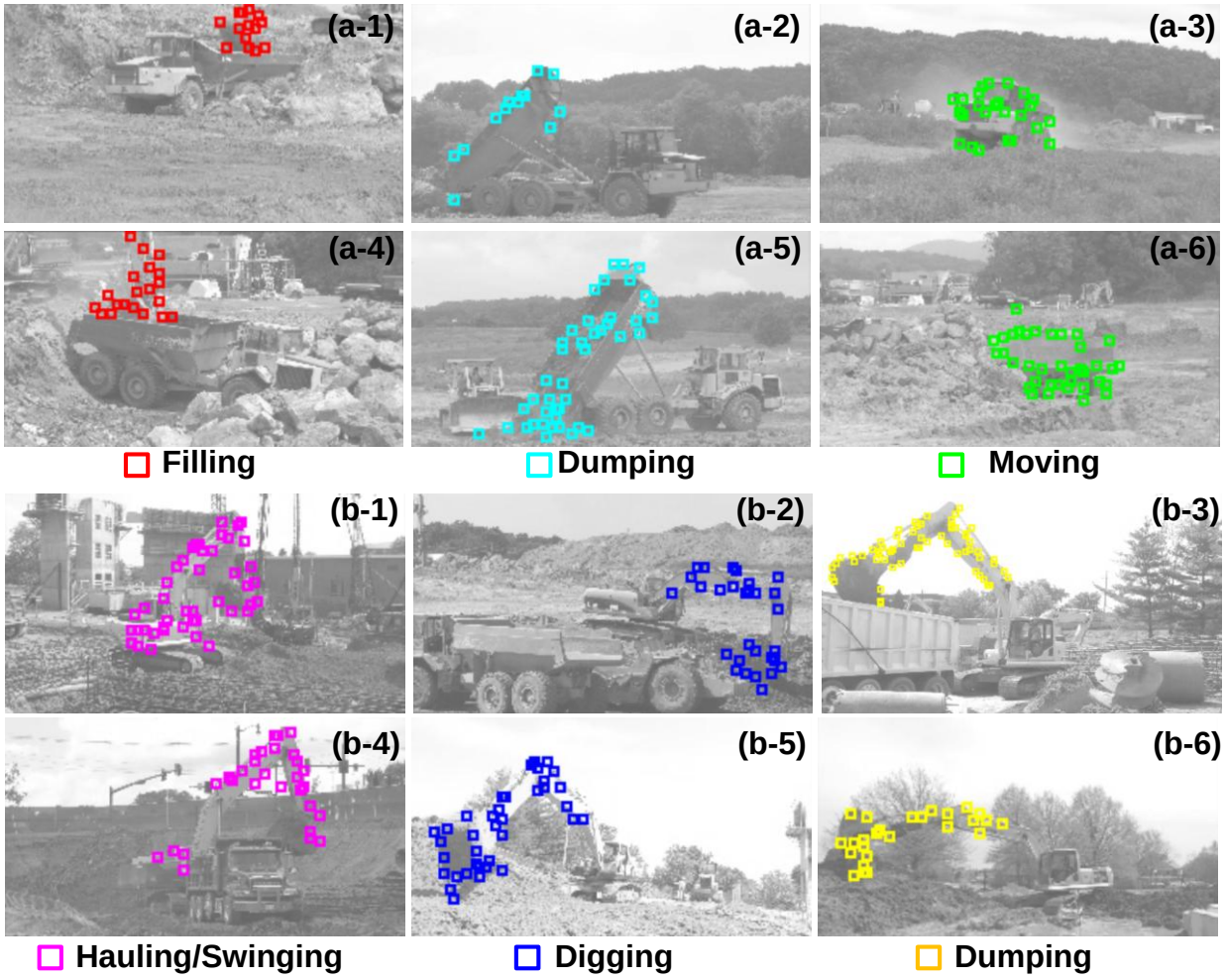


Figure 3.13: Example features from testing sequences in both truck and excavator datasets. The spatio-temporal patches in each sequence are automatically color coded according to the action classification (Figure best seen in color). (a:4-6) and (b:1–3) are showing the presence of occlusions in the dataset.

3.4.4 Discussion on Model Parameters

In the following subsections, we test the effect of the feature detection parameters, the type and size of the feature descriptor, the codebook sizes, and various machine learning algorithms on the average performance of the action classification. The best parameters are selected based on reasonable accuracy and computational times, and were presented in section 3.4.2.

a. Feature Detection Parameters

The effect of the two parameters σ and τ which correspond to the spatial and temporal scales of the detectors were tested under different parameters for the excavators' actions. Figure 3.14 shows that the combination of σ = 1.5 and τ =3 which the HOG descriptors, codebooks of size 350, and the multi-class SVM classifier results in the highest average accuracy. This means that the temporal scales of features can have a higher influence on average action classification accuracy.

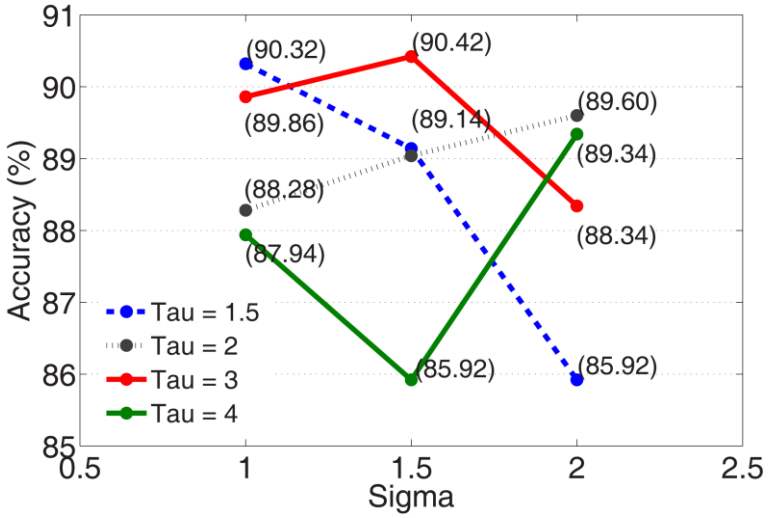


Figure 3.14: Excavator action classification accuracy vs. σ and τ feature detection values. σ=1.5 and τ=3 provides the highest accuracy of 90.42%.

b. Type of Feature Descriptor

Two types of feature descriptors: 1) HOG; 2) Histograms of Optical Flow (HOF) were tested to determine which one results in a better performance. As illustrated in Figure 3.15, while considering the codebook size to be 350, σ = 1.5 and τ =3, the HOG descriptors with 180 bins show better performance in comparison to HOF descriptors with 198 bins.

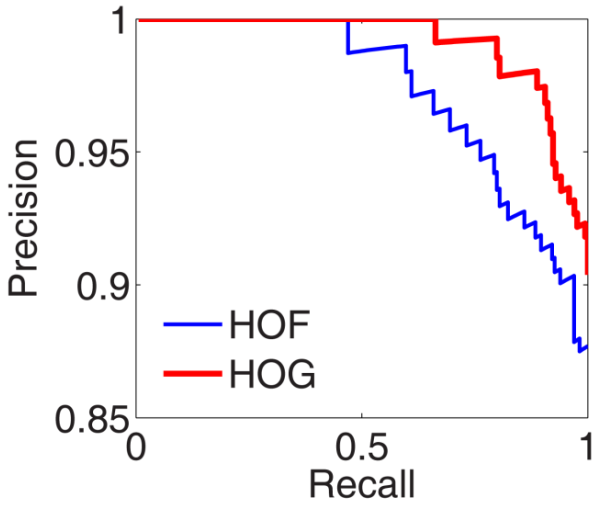


Figure 3.15: Classification precision-recall using HOG and HOF descriptors for excavator action classification.

c. Codebook Size and Histogram Formation

As explained in 3.3.2 through the *k*-means algorithm, the distance of each descriptor to each codebook center was computed, and a codebook membership was given to each HOG descriptor. To generate the best codebooks, the effect of the codebook size on the average accuracy of the multi-class binary SVM kernels is also studied. Figure 3.16 shows the classification accuracy vs. the codebook size for the excavator video datasets. As observed, codebook histograms with the size of 350 code words result in the highest action classification accuracy. While for the case of 600 bins, a similar level of accuracy is observed, nonetheless to minimize the computation time, the codebook with 350 codewords is selected.

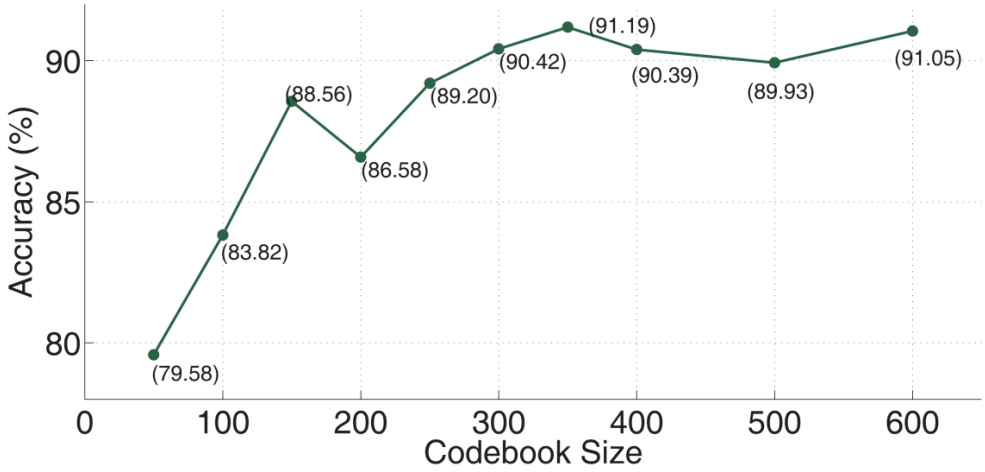


Figure 3.16: Classification accuracy obtained on the excavator video dataset using the multiple binary SVM classifiers vs. codebook size. The codebook size of 350 provides the highest accuracy of 91.19%.

d. Machine Learning Component

We have also studied the impact of using different supervised and unsupervised machine learning algorithms. Particularly the multiple linear SVM proposed in our algorithm is compared with Naïve Bayes and pLSA unsupervised algorithms proposed in (Gong et al. 2011). As observed in Figure 3.17, the performance of the multiple linear SVM is superior to the competing algorithms. This is consistent with our intuition that in the case of construction equipment and their actions where intra-class variably is significant, the supervised SVM classifier algorithm should perform better.

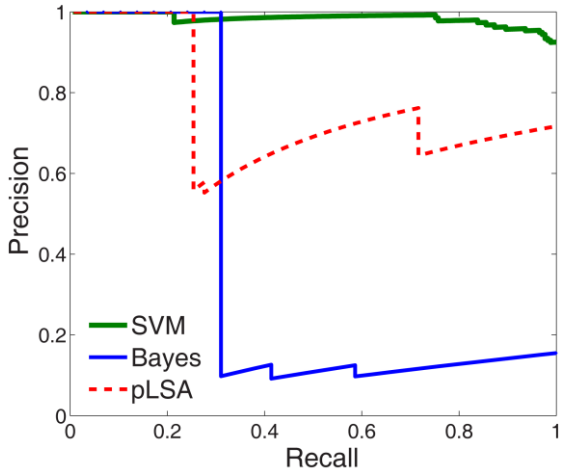


Figure 3.17: Classification precision-recall curves generated using multiple linear SVM, Naïve Bayes, and pLSA classifier algorithms.

3.5 Discussion on the Proposed Method and Research Challenges

This study presented the first comprehensive video dataset for action recognition of excavator vs. dump truck in earthmoving operations. The average accuracy of the action classification obtained for both excavator and dump truck video datasets is 86.33% and 98.33% respectively. This performance is comparative to the state-of-the-art in both computer vision and AEC communities (Gong et al. 2011; Niebles et al. 2008). The presented results show the robustness of the proposed method to dynamic changes of illumination, viewpoint, camera resolution, and scale changes as well as static and dynamic occlusions. The minimal spatial resolution of the equipment in the videos in the range of (~80-190)×(~80-190) pixels per equipment, promises the applicability of the proposed method for existing site video cameras.

While this paper presented some initial steps towards processing site video streams for the purpose of action recognition, several critical challenges remain. Some of the open research problems for our community include:

- **Action recognition in long video sequences.** Recognizing equipment actions in long sequences of video is a difficult task as 1) the duration of actions are not pre-determined, and 2) the starting point of actions are unknown. The action recognition algorithm presented in this paper is only capable of accurately recognizing actions when the starting point and duration of each action is known as *priori* knowledge. To automatically and accurately recognize the starting point and the duration of each equipment action, more work is needed on the temporal detection of each action's starting points and duration with reasonable accuracy.

- **Multiple equipment tracking and localization.** Action recognition for multiple equipment requires precise 2D tracking and localization of equipment in the video streams. Robust tracking could also enable automated detection and 3D localization of equipment for proximity analysis purposes. It further enables the action recognition to be limited to certain regions in the video streams, further minimizing the effect of noise caused by 1) lateral movement of the camera, 2)

dynamic motions of foreground (e.g., grass or vegetation) or background (e.g., offsite pedestrians or moving vehicles), and finally 3) spatio-temporal features detected around the moving shadow of the working equipment.

- **Variability in equipment types and models.** Accuracy of action recognition is an important concern for applications such as equipment productivity or carbon footprint assessment. As a result, comprehensive dataset of all types and models of equipment from all possible viewpoints is required for model training purposes. The dataset presented in this work only includes two types of equipment from six different manufacturers. Development of larger datasets is still needed.

- **Detection of idle times.** In this paper, it is assumed that the idle times can be easily distinguished in cases where no spatio-temporal features are detected or there are detected in low numbers. Given typical non-working short time periods between equipment actions and possible noise in site video streams, it is important to conduct further studies to investigate the reasonable time periods and minimal spatio-temporal features that can be considered as idle times.

3.6 Conclusion

This paper presents a new method for automated action recognition of earthmoving equipment from a network of fixed video cameras. The experimental results with average accuracies of 86.33% and 98.33% for excavator and truck action recognition respectively hold the promise for applicability of the proposed method for automated construction activity analysis. The robustness of the proposed approach to variations in size and type of construction equipment, camera configuration, lighting condition or presence of occlusions further strengthens the proposed method. Compared to other sensing technologies (e.g., GPS, wireless trackers), the application of video cameras is practical as it does not require additional hardware for tagging construction entities and more importantly can recognize actions. Successful execution of the proposed research will transform the way construction operations are currently being monitored. Construction operations will be more frequently assessed through an

inexpensive and easy to install solution, thus relieving construction companies from the time-consuming and subjective task of manual method analysis of construction operation or installation of expensive location tracking and telematics devices.

The current model is capable of automatically recognizing the actions of the construction equipment for a given video captured from all possible viewpoints, scales, and illuminations. In order to provide a comprehensive method for automated productivity and emission analysis, future work will include action recognition in long video sequences, multiple equipment tracking and localization, detection of idle times, and improving the dataset for better consideration of possible variability in equipment type and model. As part of a larger research project, these are currently being explored.

3.7 Acknowledgements

The authors would like to thank the Virginia Tech Department of Planning, Design and Construction, Holder and Skanska construction companies for providing access to their jobsites for a comprehensive data collection. The support of RAAMAC lab's current and former members, Chris Bowling and David Cline, Hooman Rouhi, Hesham Barazi, Daniel Vaca, Marty Johnson, Nour Dabboussi, and Moshe Zelkowicz is also appreciated. The work is supported by a grant from Institute of Critical Technologies and Applied Science at Virginia Tech.

3.8 References

B. Yao, A. K., and Fei-Fei, L. (2011). "Classifying Actions and Measuring Action Similarity by Modeling the Mutual Context of Objects and Human Poses." *International Conference on Machine Learning (ICML)*.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. "Actions as space-time shapes." *Proc., Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 1395-1402 Vol. 1392.

Brilakis, I., Park, M., and Jog, G. (2011). "Automated vision tracking of project related entities." *Advanced Engineering Informatics*, 25(4), 713-724.

Burges, C. J. C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Min. Knowl. Discov.*, 2(2), 121-167.

Chang, C.-C., and Lin, C.-J. (2011). "LIBSVM : a library for support vector machines. ." *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27.

Cheng, T., Venugopal, M., Teizer, J., and Vela, P. A. (2011). "Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments." *Automation in Construction*, 20(8), 1173-1184.

Cheung, V., Frey, B. J., and Jojic, N. "Video epitomes." *Proc., Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 42-49 vol. 41.

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

CII (2010). "Leveraging Technology to Improve Construction Productivity, Volume III: Technology Field Trials." RR240-13.

Dalal, N., and Triggs, B. "Histograms of oriented gradients for human detection." *Proc., Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 886-893 vol. 881.

Dalal, N., Triggs, B., and Schmid, C. (2006). "Human Detection Using Oriented Histograms of Flow and Appearance

Computer Vision – ECCV 2006." A. Leonardis, H. Bischof, and A. Pinz, eds., Springer Berlin / Heidelberg, 428-441.

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. "Behavior recognition via sparse spatio-temporal features." *Proc., Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, 65-72.

Efros, A. A., Berg, A. C., Mori, G., and Malik, J. "Recognizing action at a distance." *Proc., Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 726-733 vol.722.

El-Omari, S., and Moselhi, O. (2009). "Data acquisition from construction sites for tracking purposes." *Engineering, Construction and Architectural Management*, 16(5), 490 - 503.

EPA (2010). "Climate change indicators in the united states." *USEPA*, EPA 430-R-10-00. .

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

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). "Object Detection with Discriminatively Trained Part-Based Models." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9), 1627-1645.

Feng, X., and Perona, P. "Human action recognition by sequence of movelet codewords." *Proc., 3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*, 717-721.

Frey, C., Rasdorf, W., and Lewis, P. (2010). "Comprehensive Field Study of Fuel Use and Emissions of Nonroad Diesel Construction Equipment." *Journal of the Transportation Research Board*, 2158, 69-76.

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

Gong, J., and Caldas, C. H. (2008). "Data processing for real-time construction site spatial modeling." *Automation in Construction*, 17(5), 526-535.

Gong, J., and Caldas, C. H. "An Intelligent Video Computing Method for Automated Productivity Analysis of Cyclic Construction Operations." ASCE, 7-7.

Gong, J., and Caldas, C. H. (2010). "Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations." *Journal of Computing in Civil Engineering*, 24(3), 252-263.

Gong, J., Caldas, C. H., and Gordon, C. (2011). "Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models." *Advanced Engineering Informatics*, 25(4), 771-782.

Goodrum, P. M., Haas, C. T., Caldas, C., Zhai, D., Yeiser, J., and Homm, D. (2011). "Model to Predict the Impact of a Technology on Construction Productivity." *Journal of Construction Engineering and Management*, 137(9), 678-688.

Grau, D., and Caldas, C. H. (2009). "Methodology for Automating the Identification and Localization of Construction Components on Industrial Projects." *Journal of Computing in Civil Engineering*, 23(1), 3-13.

Grau, D., Caldas, C. H., Haas, C. T., Goodrum, P. M., and Gong, J. (2009). "Assessing the impact of materials tracking technologies on construction craft productivity." *Automation in Construction*, 18(7), 903-911.

Heydarian, A., and Golparvar-Fard, M. "A Visual Monitoring Framework for Integrated Productivity and Carbon Footprint Control of Construction Operations." ASCE, 62-62.

Heydarian, A., Golparvar-Fard, M., and Niebles, J. C. (2012). "Automated visual recognition of construction equipment actions using spatio-temporal features and multiple binary support vector machines." *Construction Research Congress*.

Hofmann, T. (1999). "Probabilistic latent semantic indexing." *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, Berkeley, California, United States, 50-57.

İkizler, N., and Forsyth, D. (2008). "Searching for Complex Human Activities with No Visual Examples." *International Journal of Computer Vision*, 80(3), 337-357.

Laptev, I. (2005). "On Space-Time Interest Points." *International Journal of Computer Vision*, 64(2), 107-123.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. "Learning realistic human actions from movies." *Proc., Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1-8.

Laxton, B., Jongwoo, L., and Kriegman, D. "Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video." *Proc., Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 1-8.

Lewis, P., Leming, M., Frey, C., and Rasdorf, W. (2011). "Assessing the Effects of Operational Efficiency on Pollutant Emissions of Nonroad Diesel Construction Equipment." *Journal of the Transportation Research Board*, 11-18.

Liu, J., and Shah, M. "Learning human actions via information maximization." *Proc., Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1-8.

Marszalek, M., Laptev, I., and Schmid, C. "Actions in context." *Proc., Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2929-2936.

Navon, R., and Sacks, R. (2007). "Assessing research issues in Automated Project Performance Control (APPC)." *Automation in Construction*, 16(4), 474-484.

Niebles, J., Wang, H., and Fei-Fei, L. (2008). "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words." *International Journal of Computer Vision*, 79(3), 299-318.

Oglesby, C. H., Parker, H. W., and Howell, G. A. (1989). "Productivity Improvement in Construction." *McGraw-Hill, New York, NY* 84-130.

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

Rezazadeh Azar, E., and McCabe, B. (2011). "Automated Visual Recognition of Dump Trucks in Construction Videos." *Journal of Computing in Civil Engineering*, In Press.

Rish, I. (2001). "An empirical study of the naive Bayes classifier " *International Joint Conf. on Artificial Intelligence*.

Savarese, S., DelPozo, A., Niebles, J. C., and Fei-Fei, L. "Spatial-Temporal correlatons for unsupervised action classification." *Proc., Motion and video Computing, 2008. WMVC 2008. IEEE Workshop on*, 1-8.

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

Song, J., Haas, C. T., and Caldas, C. H. (2006). "Tracking the Location of Materials on Construction Job Sites." *Journal of Construction Engineering and Management*, 132(9), 911-918.

Su, Y. Y., and Liu, L. Y. "Real-Time Construction Operation Tracking from Resource Positions." ASCE, 25.

Teizer, J., Lao, D., and Sofer, M. (2007). "Rapid Automated Monitoring Of Construction Site Activities Using Ultra-Wideband." *The 24th International Symposium on Automation and Robotics in Construction. ISARC 2007*, Published by I.A.A.R.C., p.23-28.

Vapnik, V., and Bottou, L. (1977). "On structural risk mnimization or overall risk in a problem of pattern recognition " *Automation and Remote Control*.

Wang, Y., and Mori, G. (2009). "Human Action Recognition by Semilatent Topic Models." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(10), 1762-1774.

Wang, Y., Tran, D., and Liao, Z. "Learning hierarchical poselets for human parsing." *Proc., Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1705-1712.

Williams, C., Cho, Y. K., and Youn, J.-H. "Wireless Sensor-Driven Intelligent Navigation Method for Mobile Robot Applications in Construction." ASCE, 76-76.

Wong, S.-F., Kim, T.-K., and Cipolla, R. "Learning Motion Categories using both Semantic and Structural Information." *Proc., Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 1-6.

Yang, J., Vela, P. A., Teizer, J., and Shi, Z. K. "Vision-Based Crane Tracking for Understanding Construction Activity." ASCE, 32-32.

Yang, Y., and Ramanan, D. "Articulated pose estimation with flexible mixtures-of-parts." *Proc., Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1385-1392.

Yao, B., and Zhu, S.-C. (2009). "Learning deformable action templates from cluttered videos." *International Conference On Computer Vision (ICCV)*, 1-8.

Zhai, D., Goodrum, P. M., Haas, C. T., and Caldas, C. H. (2009). "Relationship between Automation and Integration of Construction Information Systems and Labor Productivity." *Journal of Construction Engineering and Management*, 135(8), 746-753.

Zou, J., and Kim, H. (2007). "Using Hue, Saturation, and Value Color Space for Hydraulic Excavator Idle Time Analysis." *Journal of Computing in Civil Engineering*, 21(4), 238-246.

Chapter 4: Conclusion and Future Works

4.1 Summary

Over the past few years, cheap and high resolution digital cameras, extensive data storage capacities, in addition to the availability of internet on construction sites have enabled capturing and sharing of construction video streams on a truly massive scale. Detailed and dependent video streams provide a transformative potential of gradually and inexpensively sensing action and location of construction equipment, enabling construction firms to remotely analyze progress, safety, quality, productivity, and carbon footprint. Using a network of high definition video cameras, in this thesis, a new method for automated 2D detection, 3D tracking, and action recognition of construction equipment is presented in this thesis. These methods are briefly described in Chapter 1 along with research objectives, methodology, and thesis overview.

The study in chapter 2 presents a computer vision based algorithm for automated 2D tracking and 3D localization of construction equipment from site video streams. The state-of-the-art research proposes semi-automated methods for tracking of construction equipment. Chapter 2 summarizes the methodology and the developed algorithm to automatically track and localize construction equipment. Chapter 3 summarizes the developed algorithm to automatically recognize actions of construction equipment. This research is a step towards fully automated monitoring and analysis of operational performance. Being able to automatically detect construction resources, localize them in 3D, and simultaneously recognize their actions allows project managers to improve operational performances by adjusting different sequences and creates a safer work environment for the workers and operators on site. I hope one day I will be able to see these algorithms properly software engineered and be used on construction sites for performance monitoring on daily basis, resulting in safer and more productive environments.

4.2 Contributions

This research presents a new technique for simultaneous 2D recognition, 3D tracking, and action recognition of earthmoving construction equipment from a network of fixed video cameras. By recognizing the operational sequence, an automatic productivity analysis can be performed. Compared to other sensing technologies (e.g., GPS, wireless trackers), this application is practical as it does not require "tagging" of construction entities. Considering the $900 billion construction industry, each 0.1 percent increase in efficiency can lead to $900 million in savings, resulting in a significant impact on the current construction practice.

4.2.1. Comprehensive Dataset

Due to lack of existing databases for benchmarking visual detection, tracking, and actions of construction equipment, this research has significantly improved the data collection process and a comprehensive benchmarking video dataset is created which could be used for training and testing purposes. This dataset will be released to the community for further development and validation of new algorithms and ideally be used as a benchmark for future works in this area. For this purpose, 300 hours of video streams recorded from five different construction projects (i.e., two building and three infrastructure projects) were collected. 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, from different views and scales in a video frame, several cameras were set up in two 180° semi-circles (each camera roughly 45° apart from one another) at the training stage. This strategy enables the equipment to be videotaped at two different scales (full and half high definition video frame heights). Particularly the following combinations of equipment are considered in our database: 1) excavators and dump trucks, 2) backhoes and dump trucks, 3) scrapers, excavators, and dump trucks, 4) scrapers, dozers, and dump trucks, and finally 5) loaders and dump truck. For each of these combinations, we recorded a distinct video database containing all possible actions of the equipment. For example, for the combination of excavators and dump trucks, this video database contains five types of

excavator actions (i.e., moving, digging, hauling [swing with full bucket], swinging [empty bucket], and dumping) and three types for dump truck actions (i.e., moving, filling, and dumping). This dataset contains three types of excavators (manufacturers: Caterpillar, Komatsu, and Kobelco) and three types of dump trucks (manufacturers: Caterpillar, Trex, and Volvo).

For 2D tracking and 3D localization, a total of 4175 and 3646 frames (positive and negative) were manually segmented, labeled, and used for initial training datasets of excavators and trucks respectively. The negative images for each binary classification's not-to-be-detected instances include both the other class' positive instances and an additional 500 negative frames which represent typical construction operations with vary dynamic backgrounds. This dataset will be made public at http:///www.raamac.cee.vt.edu/realtimetracking.

For automatic action recognition, overall a total of 150 to 170 training videos were annotated and temporally segmented for each action of equipment (overall 895 videos for four and three action classes of excavators and dump trucks). Each video has different durations, and hence various possible temporal scales for each action are introduced into the training dataset. The "idle" action category is not used for training purposes. Rather idle time frames are determined when there are no spatio- temporal features detected for a given number of consecutive frames. The video dataset will be made public at: (www.raamac.cee.vt.edu/equipmentactionrecognition).

4.2.2. Performance Assessment

With the new set of EPA regulations and current economy crisis, being able to benchmark and reduce construction emissions, which is responsible for 6 percent of the total U.S. industrial-related GHG emissions, using the resources available without additional cost could be extremely beneficial. Idle reduction of construction equipment has been a major focus and challenge for the EPA and the Construction Industry Institute

(CII); the search for innovative automated technologies to analyze and detect equipment actions, specifically idle times, has been a major priority.

One of the most challenging facts in construction is accurately measuring operation details. Being able to accurately and automatically measure operational details allows for improved productivity of the operations through elimination of the idle time resulting in reduction of operational carbon footprint. Automated 2D detection, 3D localization, and action recognition of equipment, has created the opportunity to monitor operational performances with a reasonable accuracy. Once the algorithms are improved to measure the operation details with higher accuracy, productivity of construction operation can be automatically learned; through integrating an inventory of carbon emissions and operational productivity, the algorithm will be automatically able to estimate the carbon emissions of the operations. This joint assessment of productivity and carbon footprint will enable project managers to study their operations automatically and revise their construction plan and operation strategies to simultaneously reduce their carbon footprint and increase/maintain the level of productivity.

4.3 Recommendations on Future Research

While this research presented some initial steps towards processing site video streams for the purpose of 2D detection, 3D tracking, and action recognition and localization, several critical challenges remain. Some of the open research problems for our community include:

4.3.1. Algorithmic Improvements

- **Real-time tracking in long video sequences.** Real-time and automated 2D tracking and localization of resources in long sequences of videos is a difficult task as like most sliding window algorithms, suffers from slow processing speed, making it unsuitable for safety proximity analysis. The 2D tracking and 3D localization algorithms presented in this paper are only capable of accurately tracking equipment in a post processing stage which limits their application for

mainly performing action recognition To accurately track construction resources in real-time, more work is needed to implement the HOG+C based sliding window algorithm using the NVIDIA CUDA framework.

- **Variability in equipment types and models and worker body postures.** Accuracy of 2D detection is an important concern for applications such as productivity or safety proximity analysis. As such a comprehensive dataset of all types and models of equipment from all possible viewpoints is required for training purposes. The dataset presented in this work only includes two types of equipment from six different manufacturers. Development of larger datasets for equipment detection is still needed. In the case of construction workers, our dataset only included standing workers. Development of bending workers is also needed.

- **Temporal reasoning for 2D detection of resources.** Given the nature of construction project, it is very natural for construction resources to leave and come back to the field of view of a fixed camera on a jobsite. Also there might be cases for which a resource is temporally fully occluded 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 tracking and localization using mobile cameras.** The ability to track construction workers and equipment from mobile cameras can open a lot of existing opportunities for context awareness of the resources on a jobsite. For example, a camera mounted on equipment can minimize the chances of accidents by eliminating the blind spots and alert the equipment 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 tracking resources using mobile cameras.

- **Action recognition in long video sequences.** Recognizing equipment actions in long sequences of video is a difficult task as 1) the duration of actions are not pre-determined, and 2) the starting point of actions are unknown. The action recognition algorithm presented in this research is only capable of accurately recognizing actions when the starting point and duration of each action is known as *priori*. To automatically and accurately recognize the starting point and the duration of each equipment action, more work is needed on the temporal detection of each action's starting points and duration with reasonable accuracy.

- **Multiple equipment tracking and localization.** Action recognition for multiple equipment requires precise 2D tracking and localization of equipment in the video streams. Robust tracking could enable automated detection and 3D localization of equipment for proximity analysis purposes. It further enables action recognition to be limited to certain regions in the video streams, further minimizing the effect of noise caused by 1) lateral movement of the camera, 2) dynamic motions of foreground (e.g., grass or vegetation) or background (e.g., offsite pedestrians or moving vehicles), and finally 3) spatio-temporal patches detected around the moving shadow of the working equipment.

- **Variability in equipment types and models.** Accuracy of action recognition is an important concern for applications such as productivity or carbon footprint assessment. Such a comprehensive dataset of all types and models of equipment from all possible viewpoints is required for training purposes. The dataset presented in this work only includes two types of equipment from six different manufacturers. Development of larger datasets is still needed.

4.3.2. Automated Performance Assessment

- **Detection of idle times.** In this research, it is assumed that the idle times can be easily distinguished in cases where no spatio-temporal features are detected. Given typical non-working short time periods between equipment actions and

possible noise in site video streams, it is important to conduct further studies to investigate the reasonable time periods and minimal spatio-temporal features that can be considered as idle times.

- **Automated Productivity Analysis.** The overall goal of automated detection, tracking, and action recognition of the construction operation is to be able to automatically analyze operational productivity and detect idle time of the operating equipment. This will allow project managers to quickly determine the problems and adjust/improve operational performances if necessary.

- **Automated Carbon Footprint Analysis.** By developing an automated productivity analysis of operations, the actions of the equipment can be linked to a database of emissions for automated carbon footprint measurements. This allows project managers to have accurate measurements of operational emissions at any given time to ensure they are meeting the enforced regulations.